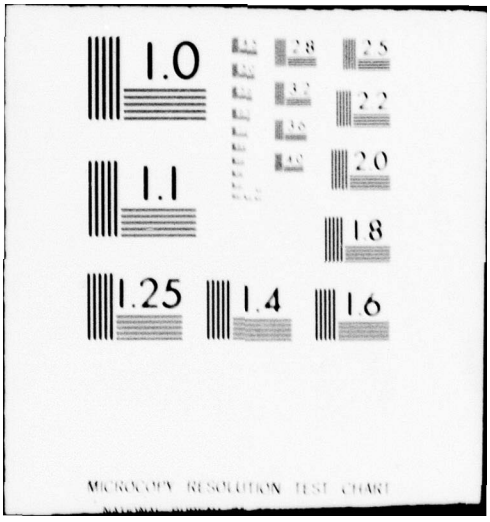
AD-A060 838 GEORGIA INST OF TECH ATLANTA SCHOOL OF ELECTRICAL EN--ETC F/G 17/2
TANDEM INTERCONNECTIONS OF LPC AND CVSD DIGITAL SPEECH CODERS.(U)
NOV 77 T P BARNWELL, R W SCHAFER, A M BUSH DCA100-76-C-0073
UNCLASSIFIED E21-685-77-TB-1 SBIE-AD-E100 092 NL

1 OF 2
AD
A060 838


1.0
1.1
1.25
1.4
1.6
2.8
2.5
3.2
2.2
3.6
2.0
4.0
1.8
MICROCOPY RESOLUTION TEST CHART

TANDEM INTERCONNECTIONS OF LPC AND CVSD DIGITAL SPEECH CODERS

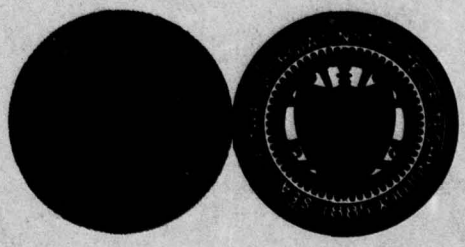
LEVEL II

By

T. P. Barnwell, III, R. W. Schafer,
and A. M. Bush

School of Electrical Engineering
GEORGIA INSTITUTE OF TECHNOLOGY
Atlanta, Georgia

FINAL REPORT E21-685-77-TB-1
Contract DCA100-76-C-0073
15 November, 1977



Prepared For
DEFENSE COMMUNICATIONS AGENCY
DEFENSE COMMUNICATIONS ENGINEERING CENTER
1860 WIEHLE AVENUE
RESTON, VA 22090

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE	READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER: E21-685-77-TB-1	2. GOVT ACCESSION NO.
3. RECIPIENT'S CATALOG NUMBER: (9) Rept.	
4. TITLE (and Subtitle): (6) Tandem Interconnections of LPC and CVSD Digital Speech Coders.	5. TYPE OF REPORT & PERIOD COVERED: Final 8 August 1976 – 30 September 1977
6. PERFORMING ORG. REPORT NUMBER: (14) E21-685-77-TB-1	
7. AUTHOR(s): (10) T. P. Barnwell, R. W. Schafer, A. M. Bush	8. CONTRACT OR GRANT NUMBER(s): (15) DCA100-76-C-0073
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Georgia Institute of Technology, School of Electrical Engineering, Atlanta, GA 30332	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: (11)
11. CONTROLLING OFFICE NAME AND ADDRESS: Defense Communication Engineering Center, 1860 Wiehle Avenue (Dr. W. R. Belfield), Reston, VA 22090	12. REPORT DATE: (15) November 1977
	13. NUMBER OF PAGES: 123
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office): (12) 134 p.	15. SECURITY CLASS. (of this report): UNCLASSIFIED
	15a. DECLASSIFICATION DOWNGRADING SCHEDULE: N/A

16. DISTRIBUTION STATEMENT (of this Report)

Unlimited, Open Publication

(18) SBIE

DISTRIBUTION STATEMENT A
Approved for public release;
Distribution Unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

Same

(19) AD-E100 992

DDC RECEIVED NOV 6 1978 B

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Speech Digitization, Linear Predictive Coder, Tandeming, PARM, Subjective Testing

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

The study described in this report was concerned with the limitations on the performance of tandem interconnections of LPC and CVSD digital speech coders. Such interconnections were systematically studied so as to identify sources of speech quality degradation and schemes for improving the performance of such tandem interconnections were investigated.

In the LPC-to-CVSD connection, the major source of speech quality degradation appears to be the minimum phase nature of the output waveform of convention-

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

408631 SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

al LPC synthesizers. The resulting "peaky" waveform causes increased slope overload distortion in the CVSD coder, thereby degrading the overall performance. To alleviate this problem, a flexible approach to modifying the phase of LPC synthetic speech was developed. Both objective and subjective tests of the phase modification system show significant but not dramatic improvements in quality for the overall LPC-to-CVSD tandem.

In the CVSD-to-LPC connection, the major source of degradation is the quantization error introduced by the CVSD coder. This quantization noise distorts the spectrum estimate obtained in the LPC analysis by broadening the bandwidths of the formant resonances. To improve the LPC spectrum estimate, the LPC analysis was modified as follows: Individual pitch periods in voiced frames were located and averaged to reduce the noise relative to the signal. A special form of periodic autocorrelation function was then computed from the averaged waveform. The resulting autoccorelation function was then processed in the normal manner to obtain the LPC parameters. This technique produced modest improvements in objective measurements of speech quality; however, no significant improvement was observed in formal subjective tests.

The implication of this study is that sophisticated code conversion systems can at best only slightly improve the quality of tandem LPC-to-CVSD and CVSD-to-LPC connections.

ABSTRACT

The study described in this report was concerned with the limitations on the performance of tandem interconnections of LPC and CVSD digital speech coders. Such interconnections were systematically studied so as to identify sources of speech quality degradation and schemes for improving the performance of such tandem interconnections were investigated.

In the LPC-to-CVSD connection, the major source of speech quality degradation appears to be the minimum phase nature of the output waveform of conventional LPC synthesizers. The resulting "peaky" waveform causes increased slope overload distortion in the CVSD coder, thereby degrading the overall performance. To alleviate this problem, a flexible approach to modifying the phase of LPC synthetic speech was developed. Both objective and subjective tests of the phase modification system show significant but not dramatic improvements in quality for the overall LPC-to-CVSD tandem.

In the CVSD-to-LPC connection, the major source of degradation is the quantization error introduced by the CVSD coder. This quantization noise distorts the spectrum estimate obtained in the LPC analysis by broadening the bandwidths of the formant resonances. To improve the LPC spectrum estimate, the LPC analysis was modified as follows: Individual pitch periods in voiced frames were located and averaged to reduce the noise relative to the signal. A special form of periodic autocorrelation function was then computed from the averaged waveform. The resulting autocorrelation function was then processed in the normal manner to obtain the LPC parameters. The technique produced modest improvements in objective measurements of speech

quality; however, no significant improvement was observed in formal subjective tests.

The implication of this study is that sophisticated code conversion systems can at best only slightly improve the quality of tandem LPC-to-CVSD and CVSD-to-LPC connections.

TABLE OF CONTENTS

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF FIGURES

LIST OF FIGURES

LIST OF TABLES

PART I

Introduction and Summary of Results

The purpose of this study was to determine the limitations on performance of tandem interconnections of LPC and CVSD digital speech coders, and to consider ways to improve the performance of such interconnections. The results are of interest in planning for a large scale secure voice transmission system which will involve both wide-band (16Kbps CVSD) and narrowband (2.4Kbps LPC) digital speech coders.

This report consists of four parts. Part I serves as an introduction and summary of results of the study. Part II is concerned with simulations of the tandem connection from the narrowband system to the wideband system. Part III deals with simulations of the Tandem connection from the wideband system to the narrowband system. Part IV discusses the details of a real-time simulation of the tandem connection using DCA's Programmable Signal Processor (PSP). Readers who are interested in a brief overview of the study and its results will find Part I adequate for this purpose. Details on specific topics are given in the other sections.

I-1. General Problem Description

Figure I-1 depicts the problem of tandem interconnections as considered in this study. Figure I-1a shows the problem that arises when a talker using a narrowband system must communicate with a listener who has available only a wideband system. First, the speech utterance is coded by an LPC vocoder, resulting in a digital representation at about 2.4 kbps. This is denoted as $\underline{c}(n)$, a vector of excitation and vocal tract response data. This digital information can be transmitted over a narrowband channel, but must be converted to the wideband representation for decoding by the wideband decoder.

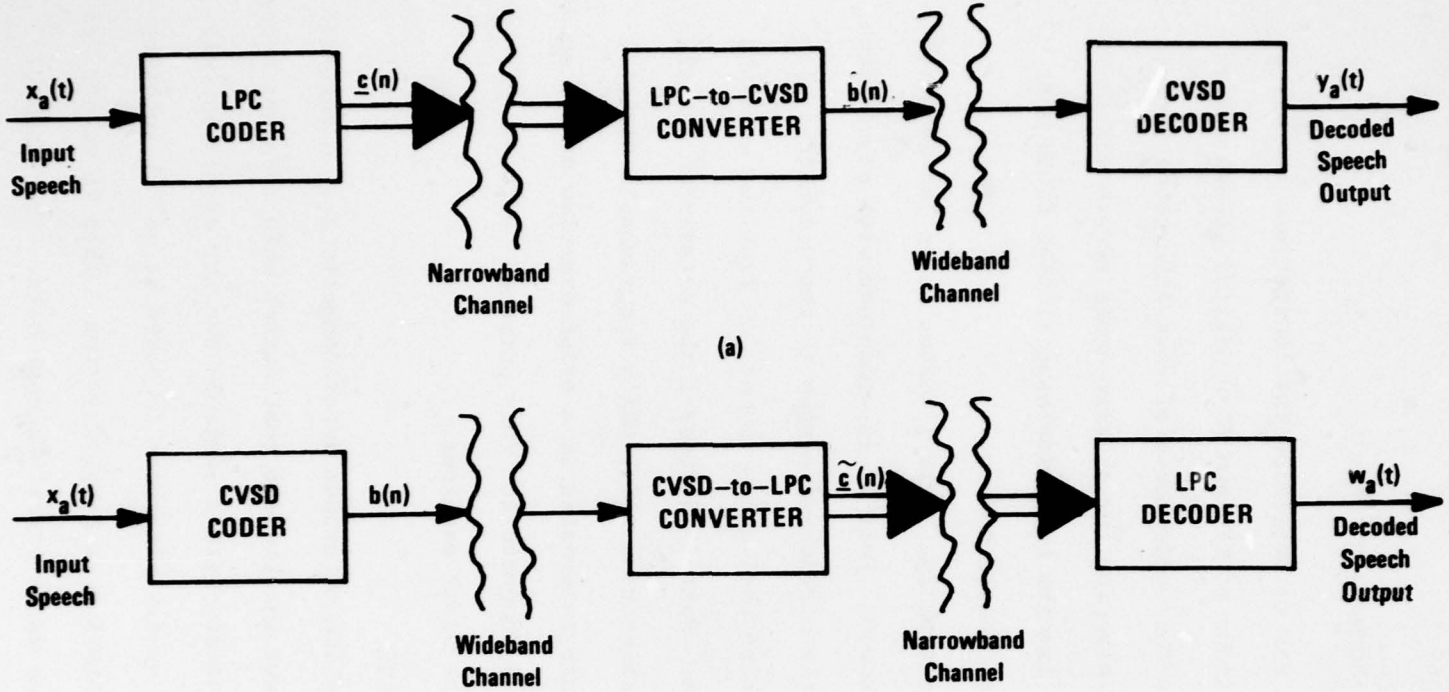


FIGURE I–1 BLOCK DIAGRAM REPRESENTATION OF TANDEM PROBLEM (a) NARROWBAND TO WIDEBAND (b) WIDEBAND TO NARROWBAND.

Thus, an LPC-to-CVSD conversion system is required for communication between the two systems. Likewise for a talker using the wideband system and a listener using a narrowband system, a CVSD-to-LPC converter is required as depicted in Figure I-1b.

A straightforward approach to this problem is to decode one representation into a speech waveform and then recode using the other system. Thus, in communication from the narrowband system to the wideband system, the LPC representation would be used to synthesize a synthetic waveform which would then be the input to a CVSD coder. The resulting binary sequence would then be transmitted in the wideband system and decoded by a standard CVSD decoder.

The problem with such tandem connections is that the conversion processes must introduce significant distortion in the overall transmission. This is evident since both LPC and CVSD by themselves introduce clearly perceptible, albeit distinctly different, types of distortion. Thus, it is apparent from the outset that the tandem connection cannot be better than its weakest component. Indeed, interactions between the two coding schemes are certain to cause additional degradation of overall performance. For example, the CVSD coding system, being designed for natural speech, under some circumstances may not perform well on the synthetic waveform produced by the LPC decoder. Likewise, the quantization noise introduced by the CVSD coding and decoding may interfere with the accurate estimation of the LPC representation.

This study therefore focused on obtaining an understanding of the degradations introduced by the conversion processes and using this knowledge, schemes for improving the overall performance were investigated.

I-2. Approach

In studying the problem of tandem interconnections of CVSD and LPC coding systems, extensive use was made of digital simulations implemented using the computer facilities of the Digital Signal Processing Laboratory at Georgia Tech [1]. These simulations are an accurate representation of typical implementations of CVSD and LPC coders, and therefore it is felt that the results of this study accurately reflect the realities of the problem. The major limitation of the simulations is that real time operation is not possible. Even so, it was possible to process six different sentences spoken by six different speakers, thus providing for a reasonably broad evaluation of the system. Real-time operation was achieved using the PSP computer at DCA/DCEC. The details of the real-time simulation are given in Part IV.

The first step in this study was to assess the degree and nature of the distortions introduced by the tandem connections. This involved both objective and subjective measures of distortion. Using the interactive computer simulation it was possible to display detailed intermediate results as well as compute average distortion measures. The information gained in this way was then used to design improved conversion systems that were compatible with the overall system constraints. Several promising ideas were evaluated on a small scale and the resultant improvements were assessed using perceptual tests. The results of these efforts are summarized in the next section.

I-3. Summary of Results

The problems of LPC-to-CVSD conversion and CVSD-to-LPC conversion are most conveniently discussed separately since they are quite distinct from each other. We begin with the problem of LPC-to-CVSD conversion.

I-3.1. Results on LPC-to-CVSD Conversion

In this case, it is obvious that the CVSD system can have no effect on the performance of the LPC system. If the CVSD system could reproduce its input waveform exactly, then the tandem connection could be no better than the LPC representation. However, this is certainly not the case for the CVSD system, and there is little hope of improving the quality of the CVSD system while retaining its basic simplicity. Thus our approach was to assume that the CVSD algorithm was fixed although its parameters could be adjusted if necessary. The design of the LPC-to-CVSD converter thus reduced to the problem of insuring that the CVSD coder could operate at peak performance on the output of an LPC synthesizer.

The major performance limitation was found to be the fact that the output waveform of an LPC synthesizer is decidedly more "peaky" than the waveform of natural speech. The concept of "peakiness" is not rigorously defined; however one possible quantitative definition is that the ratio of the peak value of a waveform to its RMS value is large. This is due primarily to the unnatural phase relationships introduced in the LPC synthesis which tend to cause large concentrations of energy at the beginning of each pitch period. This causes increased slope-overload distortion in the CVSD representation.

It was postulated that the phase is the difference between LPC synthetic speech and natural speech because the glottal excitation pulses are non-minimum phase* in natural speech, while standard LPC syn-

* In discrete-time simulations, non-minimum phase means that the z-transform of a sequence (e.g. a glottal pulse) has zeros outside the unit circle of the z-plane.

thesizers do not have the capability of producing a non-minimum phase output. Using computer simulations it was shown that for voiced speech it is possible, using a fixed second order pre-emphasis filter, to remove most of the effect of the glottal wave shape prior to LPC analysis. This permits an appropriate speech-like phase to be inserted using a non-minimum phase de-emphasis filter following conventional LPC synthesis. It was also shown that the appropriate phase can also be imposed using an all-pass filter following a conventional LPC synthesizer or by modifying the excitation of an LPC synthesizer. It was found that the output waveform of an LPC synthesizer can be made remarkably like that of natural speech.

Objective (SNR) measurements of the quality of coding achieved by the CVSD coder showed that significant improvements in signal-to-noise ratio can result from changes in the phase. These improvements are dependent upon the amplitude of the LPC waveform, being evident only for relatively large amplitudes where slope-overload effects are the major distortion.

Although improving the phase of the LPC synthetic output produces improvements in SNR, it is well known that SNR differences often do not imply significant perceptual differences. To test the perceptual effect of phase modification, a number of PARM [2] tests were run to compare the outputs of the CVSD coder under different phase and pre-emphasis conditions. The results of the tests showed that the suggested phase improvements yielded a tandem system that was slightly preferred to standard LPC synthesis; however, this preferrence was small and was not statistically significant for all cases. This is most likely due to the fact that phase improvements

are only effective in cases where slope-overload results without phase modifications. Since listeners seem to prefer slope overload distortion to granular distortion in delta modulators [3], it is not surprising that mitigating the slope-overload effect does not dramatically improve listener preference.

I-3.2. Results on CVSD-to-LPC Conversion

It is obvious as before that the system which comes last in the tandem connection (the LPC system) cannot affect the performance of the system that precedes it (the CVSD system). Again, if the CVSD system could reproduce the input with high precision, the overall quality would be that of the LPC system. Since this is not the case, it was first necessary to determine the deleterious effects of CVSD coding upon the LPC analysis.

The major effect is, of course, due to the quantization noise introduced by the CVSD system. Clearly, the LPC coder operating upon the output of a CVSD decoder is effectively representing signal plus noise. It was found that this noise interferes with the estimation of the LPC coefficients, causing the bandwidths of the vocal resonances to be greatly broadened, thereby introducing significant degradation in the LPC synthesizer output.

The basis for LPC analysis is the short-time autocorrelation function. When the input is represented as signal plus noise, a simple analysis shows that the noise enters the autocorrelations function as three terms; two crosscorrelations between the signal and noise and the short-time autocorrelation function of the noise. It was found that these terms varied greatly in size and character making fixed corrections to the autocorrelation function very impractical. This implies that improvements in tandem

performance can result only from suppressing the effects of the quantization noise prior to computation of the short-time autocorrelation function of the CVSD output. Such noise suppression is complicated by the fact that CVSD quantization noise is non-stationary and often highly correlated with the signal.

A simple approach to noise suppression was investigated in some detail. This approach involved the location of individual pitch periods in voiced frames followed by averaging to reduce the noise relative to the signal. A special form of periodic autocorrelation function was then computed as the basis for computing an LPC representation. It was found that this procedure (called the average pitch period method) produced modest improvements in the LPC spectra computed from CVSD outputs. Objective spectral distance measures showed that the average pitch period method produced modest improvements when up to three consecutive periods were averaged prior to computation of LPC parameters. Although careful listening comparisons indicate a slight improvement over conventional LPC analysis applied to the CVSD output, formal PARM tests showed no significant improvement.

I-4. Conclusions and Recommendations

The results of this study do not augur well for the use of tandem connections of LPC and CVSD coders. As we have summarized above, the deleterious interactions between the LPC and CVSD coding systems are readily understood. Their perceptual ramifications are not as easily understood but are nevertheless measurable. It appears that sophisticated conversion systems can at best slightly improve the quality of tandem LPC-to-CVSD and CVSD-to-LPC connections, and, thus, it is our judgement that this approach is not viable.

An alternative approach to the wideband/narrowband tandeming problem would be to design an improved 16 Kbps coder. Using more sophisticated processing, it is likely that quality much superior to that of CVSD could be obtained. Furthermore, a new 16 kbps coder could be designed to be compatible with both natural and synthetic speech.

PART II

LPC-to-CVSD Tandem Connection

II-1. Simulation of LPC-to-CVSD Tandem Connection

In order to study the effects of tandem connection of LPC vocoders and CVSD waveform coders, both types of systems were simulated using the Digital Signal Processing Laboratory Facility of the Georgia Tech School fo Electrical Engineering [1]. Since the properties of the tandem connection are dependent upon the nature of the LPC and CVSD coding algorithms, we shall describe the simulation in considerable detail in this section.

In defining the problem of tandem connections of LPC and CVSD coders it is useful to consider the complete communications system as composed of three parts: (a) the LPC coder, which generates a low bit-rate representation of the speech signal, (b) a system for converting from the LPC representation to a CVSD representation, and (c) a CVSD decoder for converting the CVSD bit stream to an analog signal for listening. Figure II-1 depicts the components of the computer simulation that make up these three parts of the complete system.

II-1.1. A-to-D and D-to-A Conversion

The first step in any LPC coding scheme is analog-to-digital conversion at a sampling rate which preserves sufficient bandwidth and with sufficiently fine quantization. The simulations were based upon six speech utterances (see Appendix A), which were lowpass filtered by an analog filter with cutoff frequency 3.2 kHz and then sampled at an 8 kHz sampling rate and

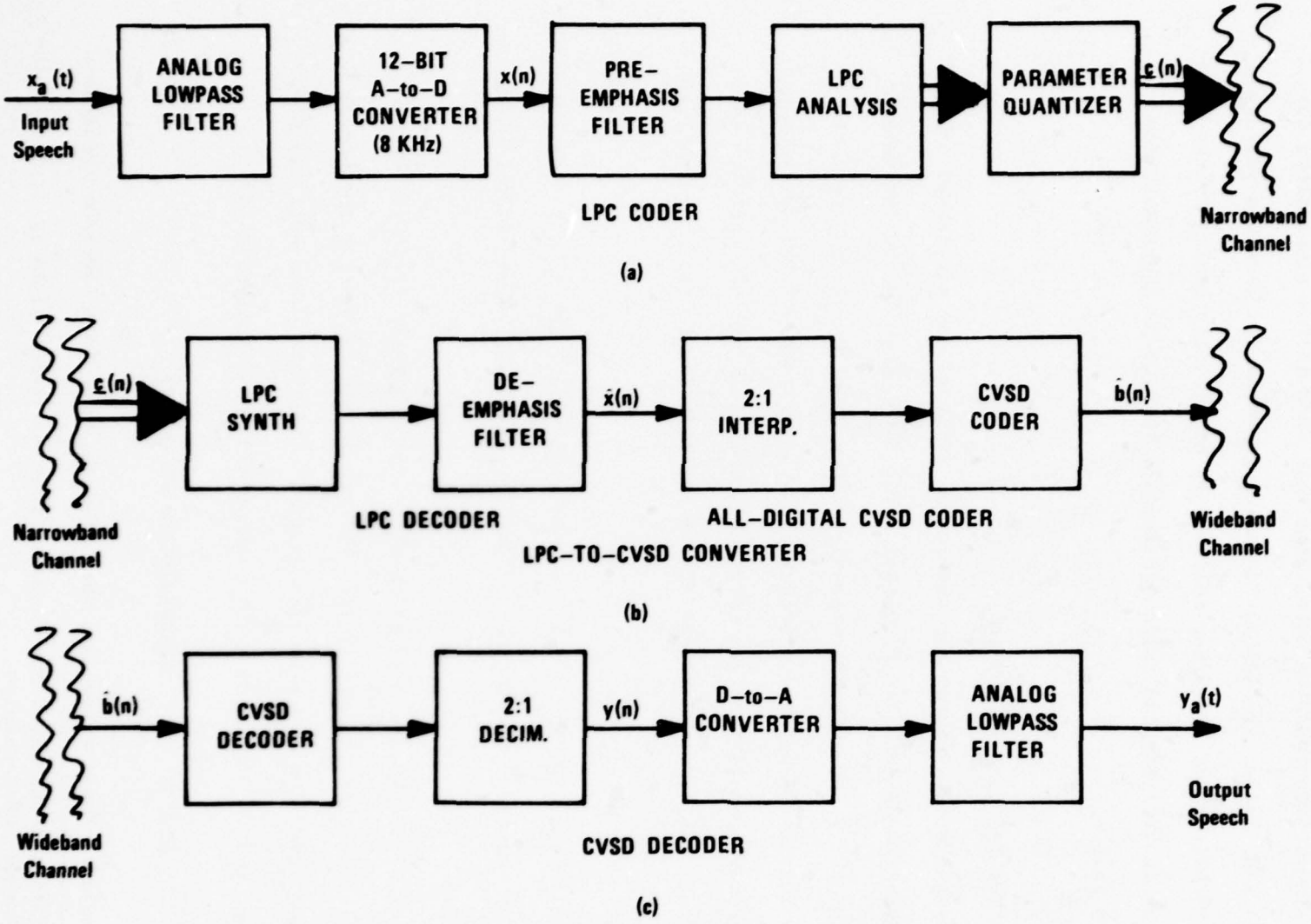


FIGURE II-1 BLOCK DIAGRAM REPRESENTATION OF LPC-to-CVSD TANDEM CONNECTION SIMULATION. (a) LPC CODER. (b) LPC-to-CVSD CONVERTER. (c) CVSD DECODER.

quantized to 12 bits per sample. These six sentences were used throughout the simulations reported here. The last step in the simulation, depicted in Fig. II-1c, is digital-to-analog conversion. This was performed using a 16-bit D-to-A converter followed by an elliptic analog filter which is flat up to 3.21 kHz and falls off rapidly to give 50 dB attenuation above 4.5 kHz.

II-1.2. The LPC Vocoder System

A complete LPC vocoder consists of an analyzer and a synthesizer. The analyzer is depicted in Fig. II-2. The first step in the implementation of an LPC vocoder is pre-emphasis of the spectrum of the input speech signal. This is necessary to reduce the dynamic range of the spectrum and thereby ease the numerical accuracy requirements in the LPC analysis [4]. For this purpose, a simple first difference filter with transfer function

$$D(z) = 1 - az^{-1} \qquad \text{(II.1)}$$

is commonly used with $.8 < a < 1$. [4] We shall see later that pre-emphasis is also an important part of our scheme for improving LPC-to-CVSD tandem performance.

After pre-emphasis, the speech signal is processed as depicted in Figure II-2. Every 15 msec. (every 120 samples) a segment of the speech waveform of duration 30 msec. (240 samples) is selected and multiplied by a Hamming window, w(n). Then the autocorrelation function values

$$R(m) = \sum_{n=0}^{L-1-m} x(n)w(n)x(n+m)w(n+m) \qquad 0 \le m \le N \qquad \text{(II.2)}$$

are computed.* These values are then used as input to a Levinson recursion

*For notational convenience, we assume that the time origin is at the beginning of the "frame" of L samples.

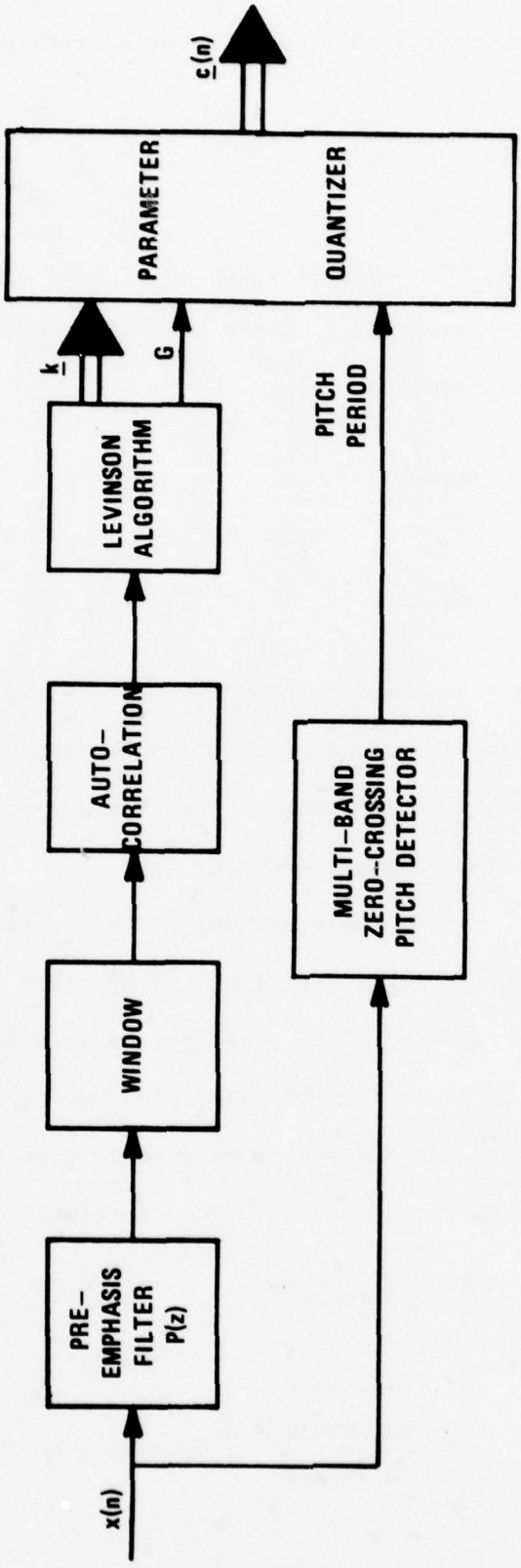


FIGURE II-2 BLOCK DIAGRAM OF LPC VOCODER ANALYZER SIMULATION.

algorithm [4,5], which produces a set of predictor coefficients

$$\underline{a} = \{a_1, a_2 \ldots, a_N\} \tag{II.3}$$

and reflection coefficients

$$\underline{k} = \{k_1, k_2, \ldots, k_N\} \tag{II.4}$$

It is well known [4,5] that knowledge of either $\underline{a}$ or $\underline{k}$ is sufficient to determine the other. Also resulting from the Levinson recursion is a gain parameter which can be expressed as

$$G = [R(0) + \sum_{j=1}^{N} a_j R(j)]$$

$$= [R(0) \prod_{j=1}^{N} (1 - k_j^2)] \tag{II.5}$$

To complete the LPC vocoder representation of the speech signal, the speech segment must be classified as either voiced or unvoiced and if voiced, the pitch period must be estimated. For this purpose, a multi-band zero-crossing pitch detector was used. [6]

For transmission over a narrowband channel it is necessary to quantize the parameters of the LPC representation. As depicted in Figure II-2, the parameters to be quantized are: the pitch period τ (note $\tau = 0$ implies unvoiced); the gain, G; and the vector of reflection coefficients, $\underline{k}$. The reflection coefficients were transformed by an inverse sine transformation and quantized with a uniform quantizer. [7] The distribution of levels is as follows:

k_1: 26 levels
k_2: 19 levels
k_3: 17 levels
k_4: 10 levels
k_5: 8 levels

k_6: 7 levels
k_7: 8 levels
k_8: 4 levels
k_9: 2 levels
k_{10}: 2 levels

The pitch period and gain were not explicitly quantized in our simulation; however, they were represented by integers. For the range of pitch periods and gain values encompassed by our set of inputs, this implies that 7 bits were effectively used to represent the pitch period and 8 bits were used for the gain. Thus, each frame required a total of 47 bits. Since the frame rate was 66.67 frames/sec., the total bit rate was 3134 bits/sec. Noting that 5-6 bits is normally considered adequate for pitch period and 3-4 bits for gain, it is reasonable to assert that comparable LPC quality could have been obtained with the system operating at 2535 bits/sec.

So far we have described the operations involved in simulating an LPC coder (Figure II-1a) as would be required for encoding speech for low bit-rate (narrowband) transmission. It is worth noting that an LPC vocoder is inherently a discrete-time system. Thus our simulation is properly viewed, not as an approximation, but as a non-real-time implementation of an LPC vocoder. The details of our simulation differ slightly from other designs, but is is in no sense an approximation to an analog system. In a real system, the quantized LPC representation would be transmitted over a communications channel, and errors would be introduced because of synchronization problems, fading, jamming, and other noise sources. Our simulation has neglected these sources of error, and therefore must be taken as an upper bound on the performance capabilities

of the LPC system and the resulting tandem system.

In communicating between a talker using a narrowband system, and one using a wideband (CVSD) system, some digital code conversion must take place. This is depicted in Fig. II-1b. It is very likely that this system would be implemented entirely using digital processing techniques even though CVSD coders normally involve a mixture of analog and digital circuitry. This being the case, the simulation of the operations depicted in Fig. II-1b may again be viewed as a non-real-time implementation of a system that could be implemented exactly in digital hardware.

The LPC representation is inherently parametric while the CVSD representation focuses on preserving the waveform of the speech signal. Therefore, to convert from LPC to CVSD requires that a waveform be reconstructed from the LPC parameters. This can be done using a standard LPC synthesizer as depicted in Fig. II-3a. In this system, an excitation signal is generated from the pitch period and gain information. If the pitch period is zero, a flat spectrum random noise generator is connected through the time-varying gain, G, to the input of a time-varying linear system. If the pitch period is non-zero, a train of unit impulses is generated with spacing equal to the pitch period. This signal is then applied through the gain, G, to the input of the system. The linear system is implemented as depicted in Fig. II-3b; i.e., as a direct form IIR digital filter [8]. The required coefficients a_1, a_2, ..., a_N are the predictor coefficients obtained in the Levinson recursion. Since only the reflection coefficients were transmitted, the quantized reflection coefficients must be converted to predictor coefficients by application of the recursion formula

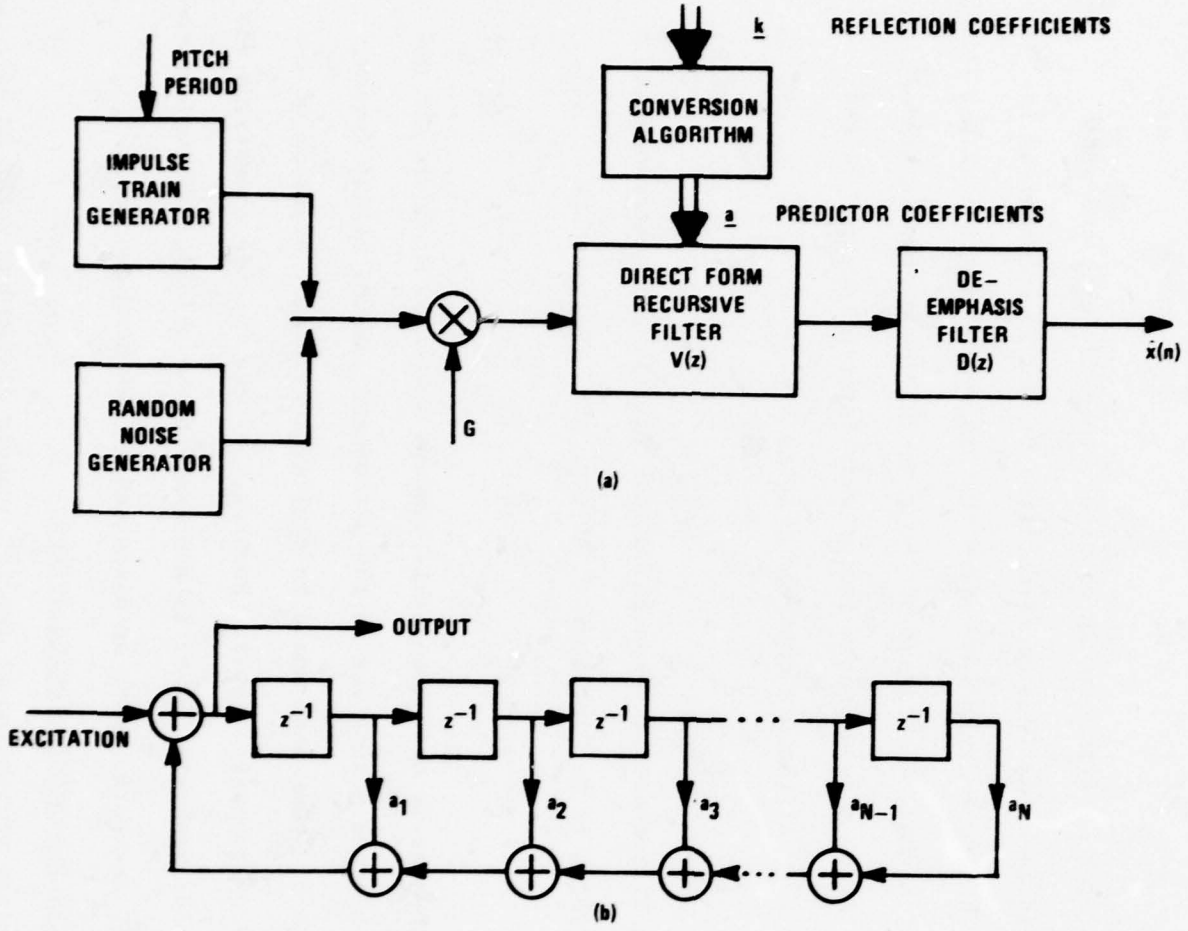


FIGURE II-3 (a) LPC VOCODER SYNTHESIZER. (b) DIRECT FORM IMPLEMENTATION OF THE SYSTEM V (z).

$$a_j^{(i)} = k_i$$

$$a_j^{(i)} = a_j^{(i-1)} + k_i \, a_{i-j}^{(i-1)} \qquad \begin{matrix} 1 \le j \le i-1 \\ 1 \le i \le N \end{matrix} \qquad \text{(II.6)}$$

where the desired predictor coefficients are

$$a_j = a_j^{(N)} \quad 1 \le j \le N. \qquad \text{(II.7)}$$

The predictor coefficients so obtained, together with the pitch period and gain, are really a representation of the pre-emphasized input speech. Thus to recover a waveform that is representative of the original input signal, the output of the time-varying synthesis filter must be filtered by a de-emphasis filter whose system function is the reciprocal of that of the pre-emphasis filter; i.e.

$$D(z) = \frac{1}{P(z)} \quad . \qquad \text{(II.8)}$$

For the first difference pre-emphasis of Eq. (II.1), the corresponding de-emphasis system is an "integrator" of the form

$$D(z) = \frac{1}{1 - az^{-1}} \qquad \text{(II.9)}$$

The LPC synthesis and de-emphasis systems make up a standard LPC decoder. When used as a receiver in the narrowband communications system, the last stage of processing would be digital-to-analog conversion of the output of the de-emphasis filter. However, for use in code conversion the output of the de-emphasis filter is processed by an all-digital implementation of a CVSD waveform coder, as discussed in the next section.

II-1.3. The CVSD Waveform Coding System

The previous discussion of the LPC vocoder system has taken us halfway through the LPC-to-CVSD code conversion process. Up to this point all

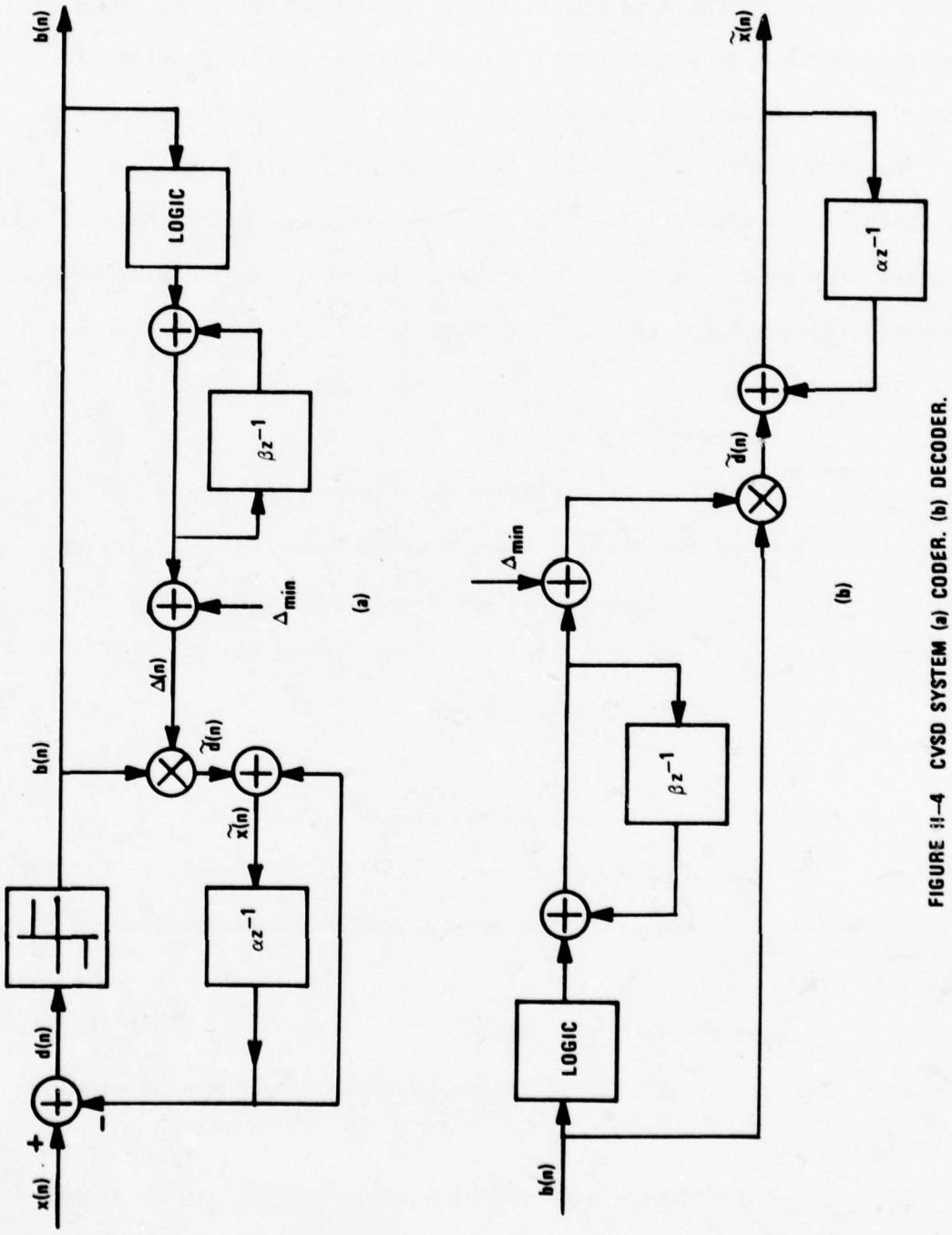


FIGURE II-4 CVSD SYSTEM (a) CODER. (b) DECODER.

speech waveforms have been sampled at an 8 kHz rate. Since the CVSD system operates at a 16 kHz sampling rate, it is necessary to interpolate the output of the de-emphasis filter to the higher rate. This is done as described in Reference [9] using a linear phase FIR digital filter. The relevant details of this process are given in Appendix B.

The CVSD coder and decoder simulations are depicted in Figures II-4a and II-4b, respectively. The CVSD coder is simply a delta modulator with adaptive step size. In Figure II-4a, $x(n)$ is the input and $\tilde{x}(n)$ is the corresponding quantized signal. The quantization is performed on a difference signal

$$d(n) = x(n) - a\,\tilde{x}(n-1) \tag{II.10}$$

which can be thought of as the difference between the input signal and a predicted value of the input. The predicted value is clearly just proportional to the previously determined quantized value. The 1-bit quantizer produces a sequence, $b(n)$, of + and -1's where

$$\begin{aligned} b(n) &= +1 \quad \text{if } d(n) \geq 0 \\ &= -1 \quad \text{if } d(n) < 0 \end{aligned} \tag{II.11}$$

The quantized difference signal is then simply

$$\tilde{d}(n) = \Delta(n) \cdot b(n) \tag{II.12}$$

where $\Delta(n)$ is the (time-varying) step-size. The quantized value of the signal is seen to be

$$\tilde{x}(n) = \alpha\tilde{x}(n-1) + \tilde{d}(n) \tag{II.13}$$

From Eqs. (II.10) and (II.13) it is easily shown that the quantization error is

$$e(n) = \tilde{d}(n) - d(n) = \tilde{x}(n) - x(n), \tag{II.14}$$

i.e., the error in the quantized signal is identical to the error in the quantized difference signal.

The step-size is varied so as to increase and decrease with the magnitude of the difference signal. By adapting the step-size from the bit stream, b(n), it is possible to derive the step-size information at the decoder without sending additional information. As seen in Fig. II-4a, the step-size is obtained from the equations

$$(\Delta(n) - \Delta_{min}) = \beta(\Delta(n-1) - \Delta_{min}) + g(n) \qquad \text{(II.15)}$$

where

$$g(n) = (1-\beta)(R-1)\Delta_{min} \quad \text{if } b(n) = b(n-1) - b(n-2)$$
$$= 0 \quad \text{otherwise} \qquad \text{(II.16)}$$

The rationale for this scheme is well known and will not be elaborated upon since our purpose is only to define the nature of our simulation.

It can be shown that for $\beta < 1$ and $g(n) = 0$, i.e., for a long run in which no three consecutive symbols are identical, then the step-size approaches Δ_{min}. Likewise; under the assumption of constant g(n) i.e. a long run of + 1's or - 1's, the maximum step-size attainable is

$$\Delta_{max} = R\,\Delta_{min} \qquad \text{(II.17)}$$

Thus, R is the ratio of maximum to minimum step-sizes.

In the LPC-to-CVSD converter, the CVSD coder could be implemented just as depicted in Fig. II-4a. However, when used in the wide-band communications system, the CVSD coding scheme would be implemented using a combination of analog and digital circuitry. For example, the first order recursive digital filters of Equations (II.13) and (II.15) would be realized using analog integrators. In the digital simulations, the signals are represented by numerical sequences while in the analog imple-

mentation, the signals are represented by voltages. In setting up the digital simulation, a correspondence was set between a voltage level of 5 volts and a numerical value of 16000. All signal levels and step-size values are set on this scale. Also, analog integrators are normally specified in terms of time constants, while the digital implementation requires the numbers α and β. The parameters of the digital simulation and corresponding analog values are given in Table II-1. Note that the range of step-sizes (R=650) is larger than normally used in CVSD coders. However, the maximum step-size is comparable to typically used values. Thus, the performance of the simulation is expected to be comparable to the analog implementation with the exception that the simulation may exhibit somewhat lower idle channel noise.

SYSTEM PARAMETER	TYPICAL VALUES FOR ANALOG IMPLEMENTATION	CORRESPONDING VALUES FOR DIGITAL IMPLEMENTATION
Maximum Signal Level	5 volts	16000
Approximate Minimum Step-Size for Best Performance	5 mv	16
Maximum Step-Size	3.34 v	10688
Predictor Time Constant	τ_a = 8 msec.	$\alpha = e^{-1/(8 \cdot 16)} = .9922$
Step-Size Filter Time Constant	τ_b = 2 msec.	$\beta = e^{-1/(2 \cdot 16)} = .9692$

Table II-1 Relationship between parameters of analog implementation and digital simulation.

The CVSD decoder is implicit in the CVSD coder since the bit stream must be decoded to obtain the predicted value of the signal used in Eq. (II.10). This is shown for completeness in Fig. II-4b. In the digital simulation, the parameters of the decoder are identical to those of the coder. In an analog implementation, exact correspondence would not, in general, be possible, and mismatch of parameters could lead to significant distortion. This source of degradation was not investigated in this study. Also, as in the case of the narrowband channel, no channel-introduced errors were simulated in the CVSD bit stream. Thus, the results of the CVSD simulation must again be viewed as an upper bound on performance.

In the simulation of the complete CVSD-to-analog system, the sampling rate of the output of the CVSD decoder was reduced from 16 kHz and then converted to an analog signal by a digital-to-analog converter and a sharp cutoff lowpass filter. The details of the 2:1 sampling rate reduction are given in Appendix B.

The performance of a digital waveform coding scheme is often judged on the basis of signal-to-quantizing noise ratio (SNR). All the signal-to-noise ratio measurements reported here were made as shown in Figure II-5. That is, the error between the input (at 8 kHz rate) and the output (at 8 kHz rate) is computed and the sum of squares was computed for all samples in the error sequence. Likewise, the sum of squares of all the samples in the test utterance was also computed. Referring to the notation of Figure II-5, the signal-to-noise ratio of the CVSD simulation is defined as

$$\text{SNR} = 10 \log_{10} \left[\frac{\sum_{\text{all } n} x^2(n)}{\sum_{\text{all } n} (\tilde{x}(n) - x(n))^2} \right] \qquad \text{(II.18)}$$

Note that no delay adjustment is required between the input and output since the interpolation and decimation filters are implemented with exactly zero phase.

As an example of performance of the CVSD system, consider Figure II-6 which shows signal-to-noise ratio as a function of signal level for each of the six sentences used in the study. In making these measurements, all the six sentences were adjusted to the same peak signal level. Variation of signal level was accomplished by varying the minimum step-size while keeping the ratio of maximum to minimum step-size the same. This is effectively the same as changing the signal level while keeping the step-size adaptation scheme fixed since the step-size is proportional to Δ_{min}. (See Eq. (II-15).) The curves of Figure II-6 (and also Figure II-18) show the variation of signal-to-noise ratio with minimum step-size on a normalized scale. The reference condition was a peak signal level of 16000 with a minimum step-size of $\Delta_{min} = 25$ and a ratio of R = 650 between maximum and minimum step-sizes. The points to the left of one correspond to smaller minimum step-sizes or equivalently larger amplitudes, and points to the right of one correspond to larger step-sizes or smaller amplitudes. Thus, the left side of Figure II-6 corresponds to the slope overload

condition while the right side corresponds to the granular noise condition. The curves of Figure II-6 display several features of the CVSD system. The maximum SNR attained varies widely among the six sentences (and six speakers); e.g. it is 8.8 dB for sentence 5 and 14 dB for sentence 3. Also, note that the location of the peak varies considerably with the utterance. A third and very important feature is the amount of variation with minimum step size for a given sentence. For example, for sentence 1, the SNR varies from 6.5 dB for large signals to 12.7 dB for the reference condition, to 7.2 dB for small signals. This strong dependence upon signal level is a severe limitation of the CVSD system and therefore of the LPC-to-CVSD tandem connection.

II-2. Performance of the LPC-to-CVSD Tandem Connection

The simulation depicted in Fig. II-1 and discussed in detail above was used to process the six sentences of Appendix A. The output of the LPC decoder was processed by the CVSD simulation for a number of different minimum step-size settings, thereby simulating different amplitude levels as in Fig. II-6. Signal-to-noise ratios were measured in each ase and compared to the signal-to-noise ratios obtained for natural speech. For most of the sentences, the natural speech gave a better signal-to-noise ratio than the LPC coded speech, although for some sentences, the LPC coded speech achieved slightly higher signal-to-noise ratios. The results of these measurements are given in Section II.3 and they are discussed in detail there.

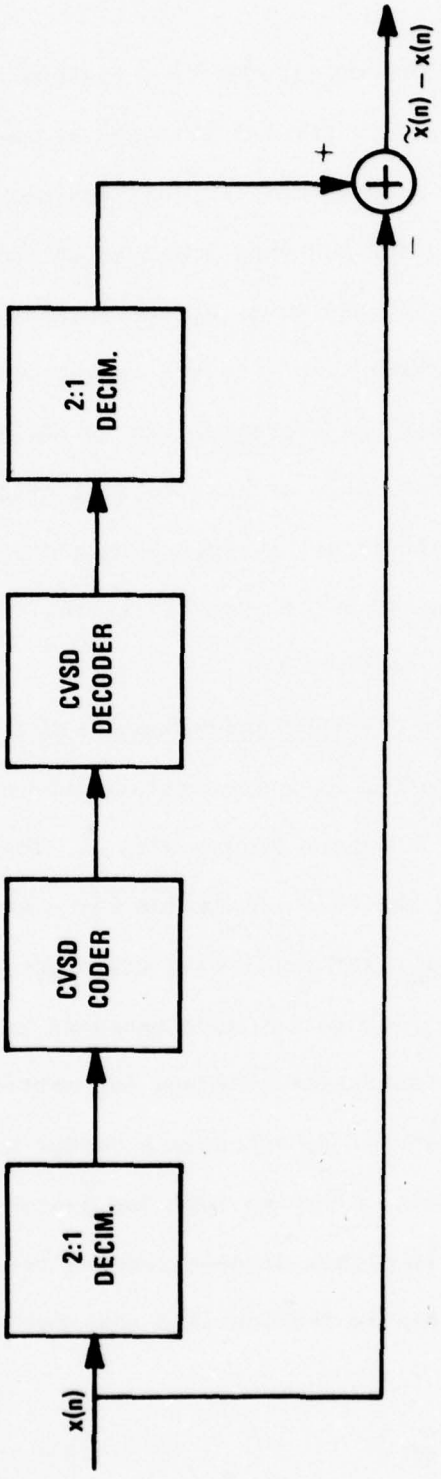


FIGURE II–5 BLOCK DIAGRAM DEFINITION OF QUANTIZATION NOISE.

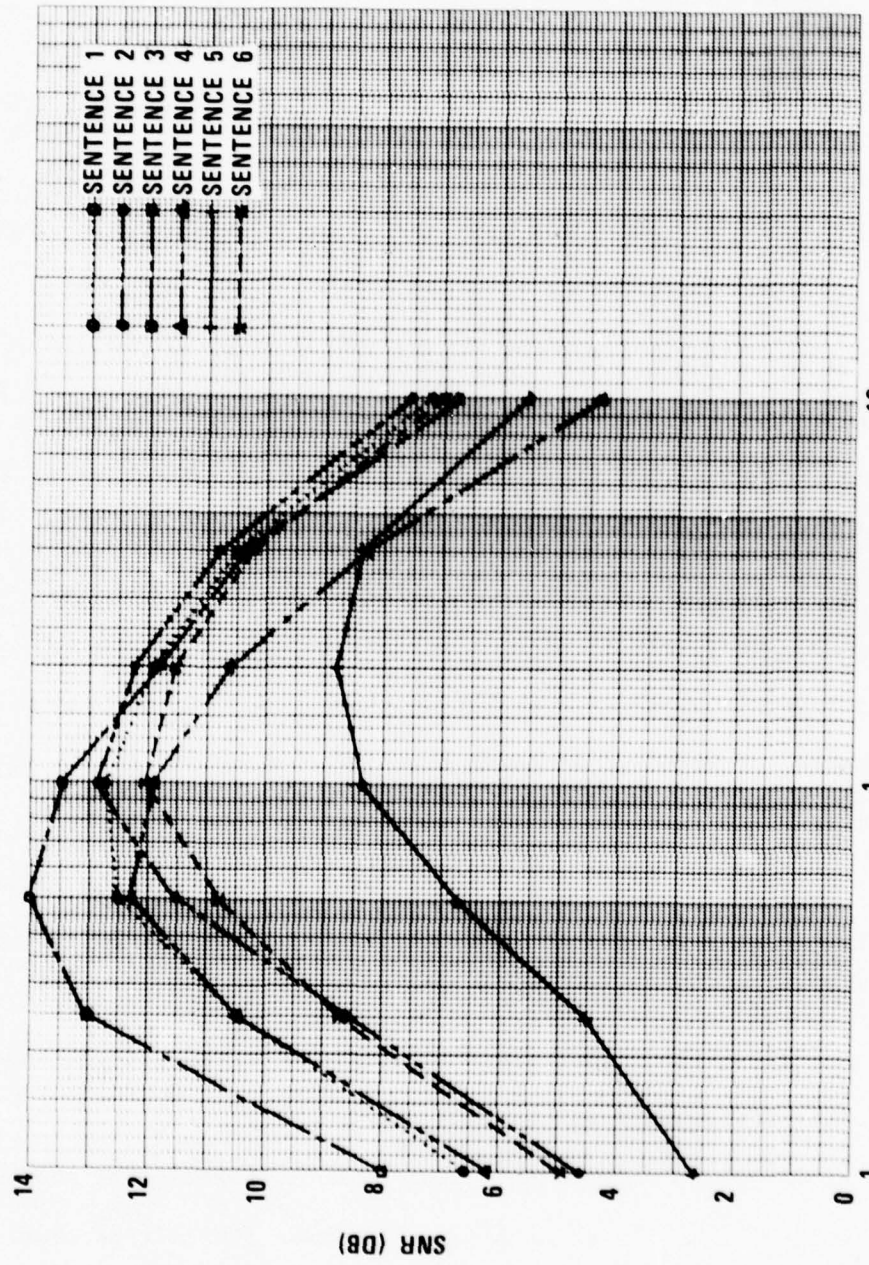


FIGURE II-6 SIGNAL-to-NOISE RATIO AS A FUNCTION OF MINIMUM STEP SIZE FOR THE CVSD SIMULATION.

The output of the LPC-to-CVSD tandem connection was also recorded on analog tape for formal and informal listening tests. These tests showed a definite perceived degradation due to the processing by both the CVSD system and the LPC system. The degradation for the CVSD system depends upon the signal level. For high signal levels, the system operates in the slope overload condition while at low signal levels, the granular quantizing noise is especially prominent. It is generally accepted that slope overload noise is preferred to granular noise [3], and this is consistent with our test results. The degradation introduced by the LPC vocoder is of a different nature from that of the CVSD system. In this case, the degradation is due to inherent limitations of the synthesis model, quantization of the model parameters, and errors in estimation of the parameters. In the LPC-to-CVSD tandem system, the degradations of both systems combine. Certainly it is unreasonable to expect the tandem system to yield a higher quality representation than either of the systems acting alone. Indeed, it might be supposed that the combination of the vocoder type distortions and waveform distortions might in some cases be significantly worse than the individual distortions. One approach to minimizing the combined effects would be to make the distortions introduced by both systems so small that the combined effects would still be acceptable. This is not possible given the rather rigid constraints on the bit-rate and structure of both the narrowband and wideband coders.

Thus, it is necessary to seek those features of the two systems that interact in deleterious ways. First of all, we can make the trivial observation that the CVSD system follows the LPC system in the LPC-to-CVSD con-

version scheme, and therefore, changes in the CVSD system cannot affect the performance of the LPC system. It is also obvious that changes in the LPC system can affect the performance of the CVSD system. Thus, we are lead to consider changes (within rather restricted bounds) of both the LPC and CVSD systems that will improve the performance of the CVSD system when operating on the output of the LPC vocoder. If the LPC vocoder were able to produce an output waveform identical to natural speech, the tandem performance could be no better than that of CVSD operating on natural speech. This suggests that one approach would be to change the LPC vocoder (or design an intermediate processor) so as to make the waveform of the LPC vocoder more like that of natural speech. This leads to the question of how the LPC vocoder output waveform differs from that of natural speech. The answer is fairly obvious from a comparison of a segment of natural speech with a corresponding segment of LPC vocoder output as in Fig. II-7. It is clear that the LPC vocoder output is much more "peaky" than the natural speech signal. Since the CVSD step-size adapts with a time constant of 2 msec., the peaks of the LPC coded speech will likely be clipped off because of slope overload. This appears to be the major difference between the waveforms of natural speech and the synthetic speech produced by the LPC vocoder. Thus, our major efforts have focused on schemes for making the LPC output waveform more like that of natural speech.

Another factor of importance is the strong dependence of CVSD performance on input signal level. This suggests that another thing that can be done is to insure that the LPC output level is always near the optimum for the setting of minimum step-sized used in the CVSD coder. This calls for an

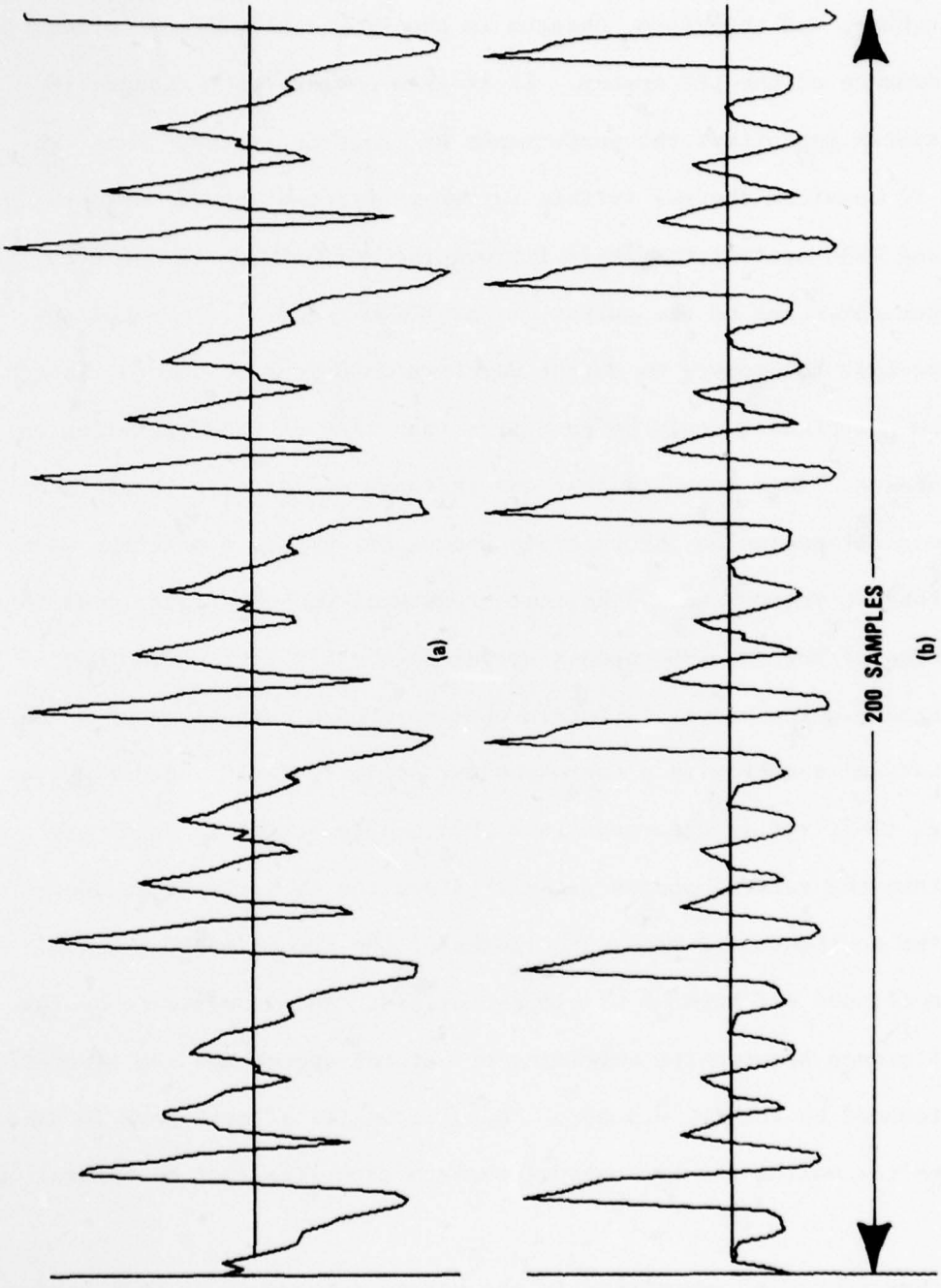


FIGURE II-7 COMPARISON OF (a) ORIGINAL SPEECH WAVEFORM AND (b) LPC SYNTHESIZED WAVEFORM.

automatic gain control to be built into the LPC vocoder synthesizer. This would be rather simple to accomplish, but to test its effect would require extensive listening in a real-time environment. Therefore this approach was not explored.

II-3. Phase Modification by Non-Minimum Phase De-Emphasis

As seen in Figure II-7, the output waveform of the LPC synthesizer is much more "peaky" than the corresponding natural speech signal from which the LPC representation was derived. The major reason for this is that the LPC synthesis system does not impart the proper phase relationships to the synthetic speech. To see this, we note from Fig. II-3 that on a short-time basis (i.e. within a frame), the transfer function of the overall synthesis system has the form

$$H(z) = V(z) \cdot D(z) \tag{II.19}$$

where

$$V(z) = \frac{1}{1 + \sum_{m=1}^{N} a_m z^{-m}} \tag{II.20}$$

and D(z) is the system function of the de-emphasis filter. It is a well known property of the LPC analysis method, that the all-pole system function, V(z), matches the magnitude spectrum of the speech signal, but not the phase. Indeed, since the poles of V(z) must lie inside the unit circle for stability, the LPC synthesizer can produce only a minimum phase output. Minimum-phase is not inherent in the production of natural speech. Although the vocal tract transmission effects can reasonably be represented by a minimum phase system, the glottal excitation pulses for voiced speech can not [10]. The effect of phase in human speech perception seems to be minimal, except for careful headphone listening [10-12]; however, it is well known that phase modifications can produce dramatic changes in wave

shape. It is reasonable to hypothesize, therefore, that a major factor in waveform differences between natural speech and LPC-coded speech is the phase response of the synthesizer.

Although the all-pole portion, V(z), of the over-all transfer function, H(z), of the synthesizer is restricted by the analysis method to be minimum phase, the de-emphasis system, D(z), still offers a possibility of inserting a non-minimum phase component into the synthetic waveform. In this section we shall discuss several approaches based on this idea.

II-3.1. The Effect of Pre-Emphasis In LPC Analysis

In seeking to modify the phase of LPC coded speech waveforms, it is reasonable to attempt to isolate the components of the spectrum which are due to the glottal excitation pulse. In the frequency domain, the major effects are a 6 - 12 dB/octave fall-off in the magnitude spectrum and the introduction of non-minimum phase components to the phase spectrum of natural speech. As was pointed out in Section II-1.1, if the LPC analysis is carried out directly on the speech signal, the fall-off of the magnitude spectrum at high frequencies may lead to ill conditioning of the LPC analysis computations. Thus, it is common practice to pre-emphasize the speech spectrum as depicted in Fig. II-2, using a first difference filter of the form

$$P(z) = 1 - az^{-1} \tag{II.21}$$

With the parameter $a \approx 1$ this filter approximately compensates for the spectrum fall-off due to the glottal pulse shape. However, better spectrum flattening can be obtained using a second order filter of the form

$$P(z) = 1 - 2r \cos \theta \, z^{-1} + r^2 z^{-2}. \tag{II.22}$$

This filter clearly has a complex conjugate pair of zeros located at a radius r and angles $\pm\ \theta$ in the z-plane. The difference between first and second order pre-emphasis is illustrated in Fig. II-8, which shows an example of the frequency response, $V(e^{j\omega T})$, and pole locations obtained with no pre-emphasis, first order pre-emphasis, and second order pre-emphasis. It is clear that second order pre-emphasis produces a much flatter spectrum and that the high frequency resonances are more clearly defined than the other two cases.

Figure II-9 shows a system for studying the effects of pre-emphasis in a systematic way. In this system, the speech is first pre-emphasized with a second order pre-emphasis filter as given in Eq. (II.22). Then a 10-pole (N = 10) LPC analysis produces the predictor coefficients required to represent V(z) as in Eq. (II.20). These coefficients were, in this case, used as the coefficients of an inverse filter for V(z); i.e.

$$V^{-1}(z) = 1+\sum_{m=1}^{N} a_m z^{-m} \qquad \text{(II.23)}$$

The output of the inverse filter should be closely related to the glottal pulse excitation in the case of voiced speech since V(z) is supposed to represent primarily the vocal-tract response. Figure II-10a shows a segment of original speech and Figure II-10b is the corresponding output of the inverse filter (labelled u(n) in Fig. II-9) when a second order pre-emphasis filter is used. It is apparent that the effects of the glottal waveshape upon the spectral magnitude can be largely removed. The waveform is very reminiscent of approximations to the glottal waveform obtained using other methods of inverse filtering [13,14]. Thus, it

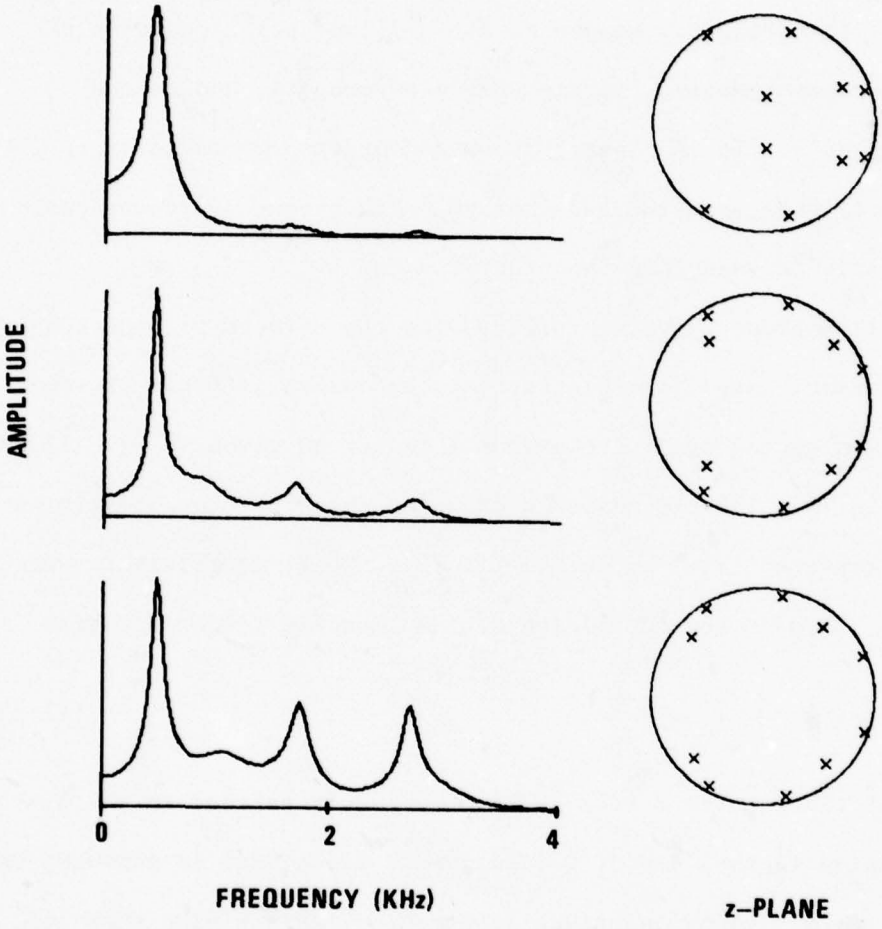


FIGURE II-8 SPECTRUM AND z-PLANE PLOT FOR: (a) NO PRE-EMPHASIS. (b) FIRST ORDER PRE-EMPHASIS (EQ (II-21). AND (c) SECOND ORDER PRE-EMPHASIS. (EQ (II-22)).

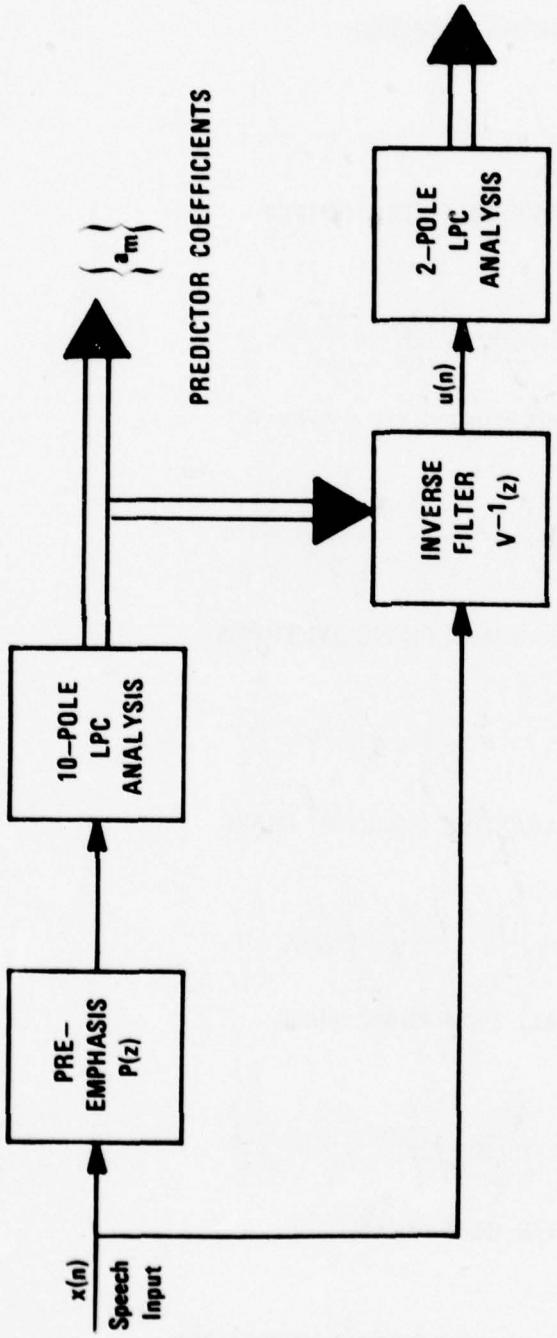


FIGURE II-9 SYSTEM FOR STUDYING THE EFFECTS OF PRE-EMPHASIS.

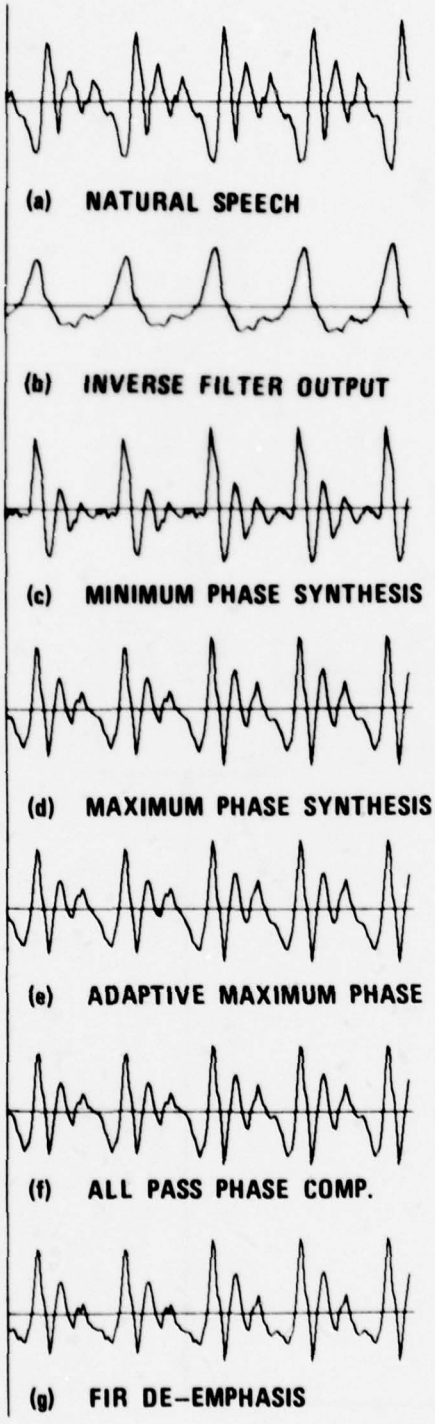


FIGURE II-10. WAVEFORM COMPARISON.

seems reasonable to assert that with proper second-order pre-emphasis, V(z) represents primarily the vocal tract response.

Given the apparent success of second order pre-emphasis in eliminating the influence of the glottal waveshape upon the LPC analysis, the question of how best to choose the parameters of Eq. (II.22) naturally arises. Another question of interest is the dependence of the pre-emphasis parameters upon speaker. To answer these questions, the following experiment was performed. The six sentences of Appendix A (note that each sentence was spoken by a different speaker) were analyzed twice using the system of Figure II-9. In the first case, there was no pre-emphasis ($P(z) = 1$ in Fig. II-9) and in the second case, second order pre-emphasis was used with parameters arbitrarily chosen to provide a reasonably good match to the properties of the glottal pulse magnitude spectrum. In each case the LPC analysis was done as discussed in Section II-1.2. For the pre-emphasized case, the speech signal was also inverse filtered as shown in Fig II-9. For each analysis frame, the poles were located and sorted by increasing angle. In addition, a two-pole LPC analysis was performed on the output of the inverse filter in the pre-emphasized case. The results of this processing are shown in the z-plane plots of Figure II-11. Figure II-11a shows the lowest frequency pole locations for voiced frames* of all six sentences for the 10-pole LPC analysis with no pre-emphasis. Note that the poles are tightly clustered; i.e. for this set of speakers and sentences, the lowest frequency pole

* Only those frames for which the gain exceeded a threshold are shown. Because the threshold was rather high, this implies that only voice frames were selected.

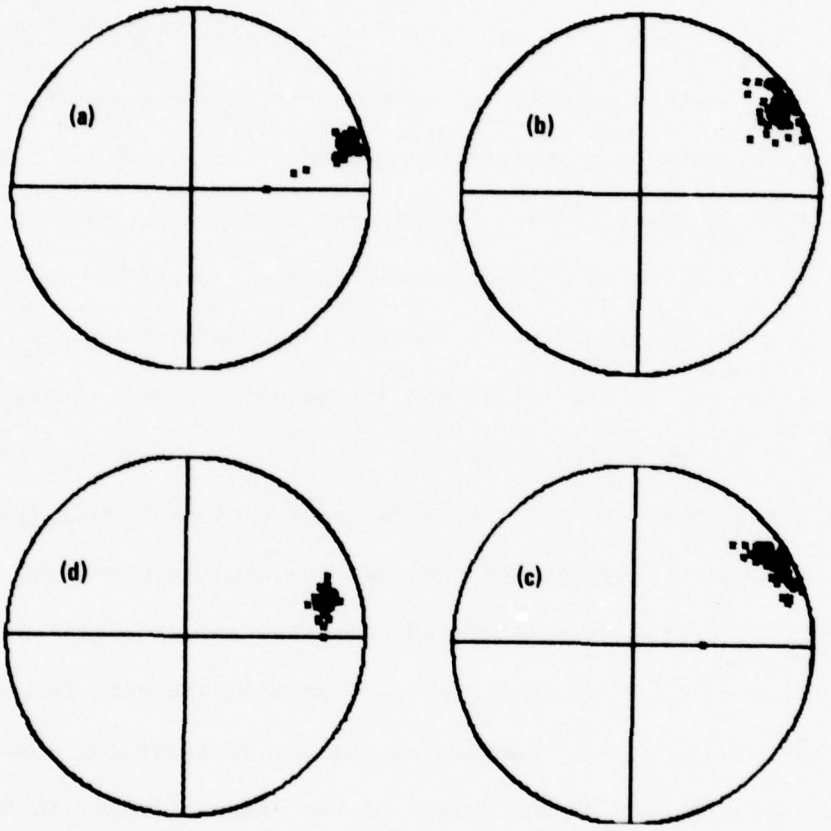


FIGURE II–11 SCATTER DIAGRAMS FOR POLE LOCATIONS OBTAINED USING FIG. II–9. (a) LOWEST FREQUENCY POLES (NO PRE–EMPHASIS). (b) SECOND LOWEST FREQUENCY POLES (NO PRE–EMPHASIS). (c) LOWEST FREQUENCY POLES (SECOND ORDER PRE–EMPHASIS). (d) TWO–POLE LPC ANALYSIS OF INVERSE FILTER OUTPUT.

varies only slightly in position from frame-to-frame. The distribution of the second lowest frequency pole in the non-pre-emphasized case, is shown in Figure II-11b. In this case the poles are more widely scattered and the center of the distribution is located at a higher frequency. The distribution of locations of the lowest frequency pole obtained with 10-pole LPC analysis of the output of the second order pre-emphasis filter is shown in Fig. II-11c. Note the similarity between Figures II-11b and II-11c. A reasonable interpretation of this is that both these distributions reflect primarily the location of the first formant frequency, while Figure II-11a must therefore be determined primarily by the glottal pulse spectrum. As a test of this hypothesis, a two-pole LPC analysis was performed on the output of the inverse filter in the pre-emphasized case. The pole locations obtained in this second order analysis are plotted in Figure II-11d. Note that the distribution is centered at low frequencies and there is very little spread. Indeed, Figure II-11d is very similar to Figure II-11a. All this suggests that the low frequency pole in the non-pre-emphasized case is closely related to the glottal components of the spectrum. Of course, when the first formant is at low frequencies, such an interpretation will not be valid, but an average of the lowest frequency pole location over a large number of frames is a reasonable starting point for obtaining a pre-emphasis/de-emphasis system.

In particular, the pole locations obtained in the two-pole analysis were averaged across the six sentences. Only those frames in which the gain determined in the LPC analysis exceed a fairly conservative threshold

were included in the average so that only voiced frames were included. The result was that the average pole location was at r = .8 and θ =.243 radians. These values were then used in the second order pre-emphasis filter, and the measurements described above were repeated. The average pole locations of the second order LPC analysis this time was r = .812 and θ = .230 radians. The close agreement of these values suggests that these would be reasonable values to use for second order pre-emphasis prior to LPC analysis.*

II-3.2. De-Emphasis Techniques for LPC Synthesis

If pre-emphasis is used in LPC analysis, then de-emphasis is required in synthesis in order to obtain proper spectral magnitude balance. The simplest approach is to set the de-emphasis filter to be the causal inverse of the pre-emphasis filter; i.e.

$$D(z) = \frac{1}{P(z)} \qquad (II.24)$$

This will restore the spectral magnitude but will again result in a minimum phase output, since the poles of D(z) (thus, the zeros of P(z)) must be inside the unit circle for stability. An example of the synthetic output for second order pre-emphasis and minimum-phase de-emphasis is shown in Fig. II-10c. Note that the waveform is quite different from the corresponding natural speech waveform in Fig. II-10a.

Since we have argued that V(z) obtained with appropriate pre-emphasis primarily represents the vocal tract, and since the excitation for voiced speech is an impulse train, then the impulse response of the de-emphasis

*Note that these values are appropriate for the sampling conditions described in Section II-1.1. Other sampling rates and filtering conditions would require different values or r and θ.

filter must represent primarily the glottal pulse. It is instructive to compare the impulse response of the de-emphasis filter to the "glottal pulses" obtained by inverse filtering. For example, if a second order complex zero (Eq. (II.22)) is used for pre-emphasis, then

$$d(n) = \left(\frac{\cos\theta n - \cos\theta(n+2)}{1 - \cos 2\theta}\right) r^n \qquad n \geq 0$$

$$= 0 \qquad n < 0 \qquad \text{(II.25)}$$

Figure II-12 shows d(n) for r = .8 and θ = .243, together with a typical "glottal pulse" derived by inverse filtering with $V^{-1}(z)$ obtained using pre-emphasis as represented by Eq. (II.22). It can be seen from Fig. II-12 that if d(n) is reversed in time, its shape is more like that of the inverse filter output. Such a de-emphasis filter has a transfer function of the form

$$D_{max}(z) = \frac{1}{P(z^{-1})} \qquad \text{(II.26)}$$

The magnitude response of this filter is identical to that of Eq. (II.24) but the phase would be the negative of the phase in Eq. (II.24); i.e. if Eq. (II.24) was minimum phase, then Eq. (II.26) would be maximum phase. Unfortunately, it is not possible to implement Eq. (II.26) as a stable and causal recursive digital filter if D(z) as given by Eq. (II.24) is stable and causal. However, for simulation purposes, the system can be implemented by applying the de-emphasis filter backwards in time; i.e. instead of time reversing the impulse response, the (finite length) input is reversed thereby producing the same effect. An example of the output of the non-minimum phase de-emphasis filter is shown in Figure II-10d. This waveform is remarkably similar to the waveform of Figure II-10a.

Even though a fixed pre-emphasis and de-emphasis was used, this "improvement" was apparent in most voiced segments throughout each of the six sentences,

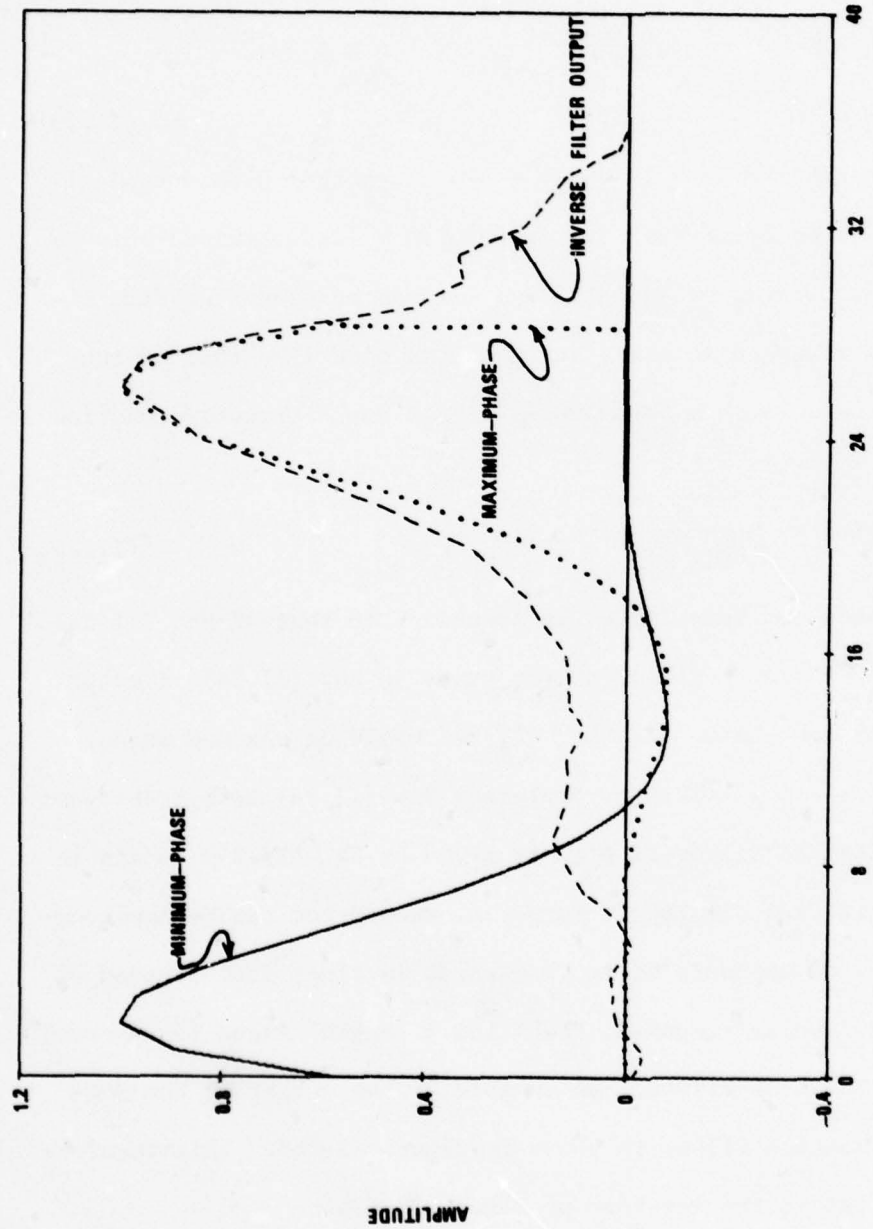


FIGURE II-12 COMPARISON OF DE-EMPHASIS FILTER IMPULSE RESPONSES AND INVERSE FILTER OUTPUT.

with little dependence upon speaker. However, it is well known that glottal waveshape can vary widely with speaker, vocal effort, and other factors. We have also seen in Fig. II-11d that LPC analysis of the inverse filtered speech shows some variability. This suggests that better performance might be obtained using a fixed pre-emphasis and adaptive de-emphasis, where the de-emphasis poles are obtained from a two-pole LPC analysis of the inverse filtered speech wave. Figure II.10e shows the resulting LPC synthesis for this case. Little difference is seen in comparing Fig. II-10e to II-10d. Indeed, little difference is noted across an utterance or between speakers. Thus, we conclude that fixed second order pre-emphasis and maximum phase de-emphasis is a valid approach to making the waveform of LPC synthesis more like that of natural speech and that adaptive de-emphasis is unnecessary. Two questions then arise. First, how can maximum phase de-emphasis be implemented in a real-time environment, and, second, does the maximum phase de-emphasis improve the LPC-to-CVSD tandem connection? The first question will be answered next and the answer to the second question is contained in Sections II-4 and II-5.

II-3.3. Implementation of Maximum Phase De-Emphasis

We have seen that the maximum-phase de-emphasis filter offers a means of imparting a more natural phase to the synthetic LPC coded speech. The time reversal filtering used in the simulation is obviously impractical; however, at least two approximate solutions are feasible.

One approach is to cascade the minimum phase de-emphasis with an all-pass filter. This is depicted in Figure II-13. Figure II-13a shows the complete LPC synthesizer with the (non-causal) maximum phase de-emphasis

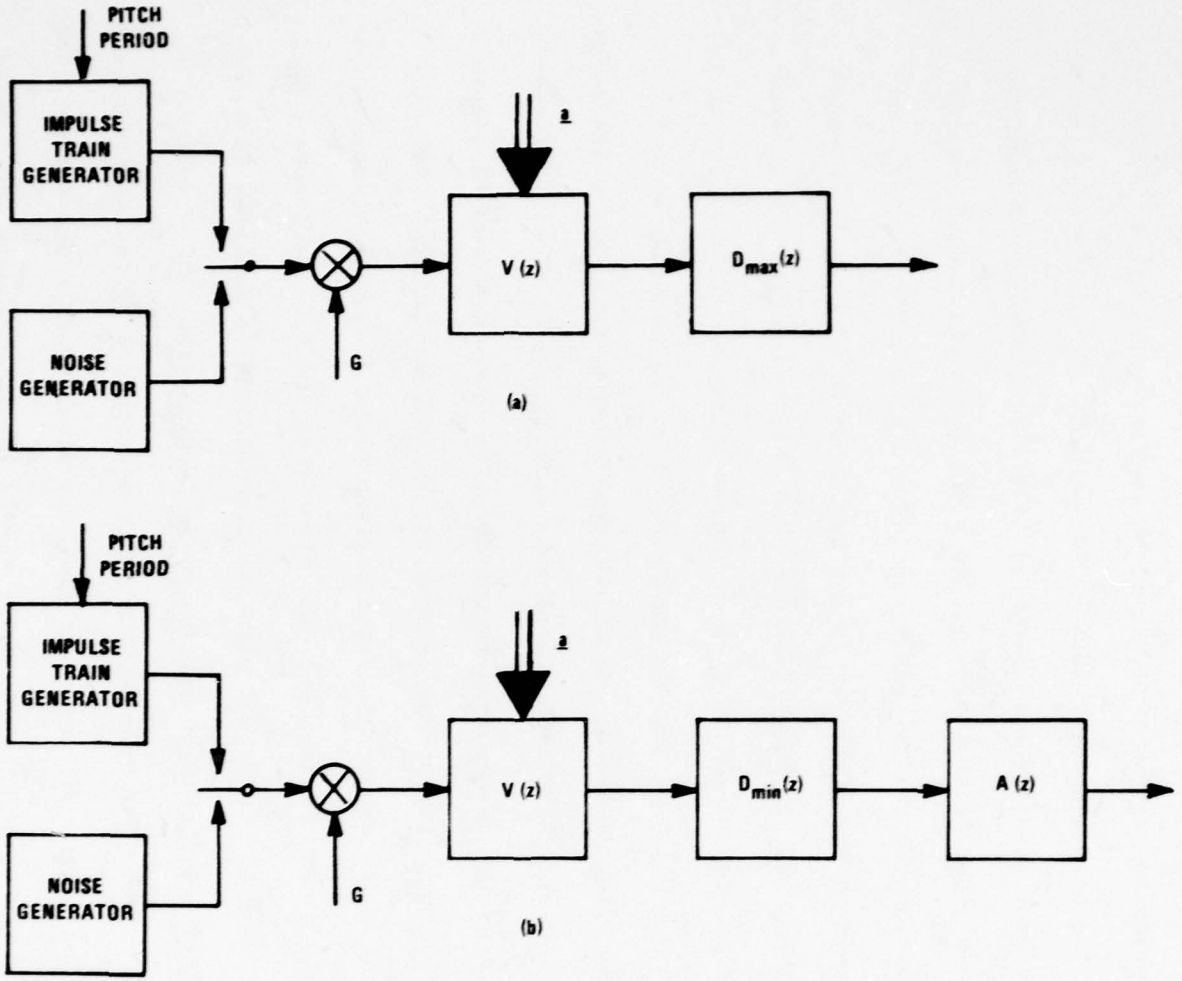


FIGURE II-13 (a) LPC SYNTHESIZER WITH MAXIMUM PHASE DE-EMPHASIS. (b) LPC SYNTHESIZER WITH MINIMUM PHASE DE-EMPHASIS AND ALL-PASS PHASE COMPENSATION.

system $D_{max}(z)$. Figure II-13b shows the LPC synthesizer in cascade with a minimum-phase de-emphasis filter

$$D_{min}(z) = \frac{1}{P(z)} \tag{II.26}$$

where the poles of $D_{min}(z)$ (zeros of $P(z)$) are inside the unit circle. Cascaded with $D_{min}(z)$ is an all-pass filter, $A(z)$, such that

$$D_{min}(z)\, A(z) = D_{max}(z) \tag{II.27}$$

The two systems of Figure II-13a and II-13b are exactly equivalent if $A(z)$ satisfies Eq. (II.27). Clearly $A(z)$ must be a non-causal system since $D_{max}(z)$ is non-causal. However, with sufficient delay, it is possible to approximate the required all-pass system very well.

The design of such an all-pass filter begins with the observation that the phase of $A(z)$ must satisfy

$$\arg[D_{max}(e^{j\omega T})] = \arg\,[D_{min}(e^{j\omega T})] + \arg\,[A(e^{j\omega T})] \tag{II.28}$$

Using Equations (II.26) and (II.27), we can write Eq. (II.28) as

$$\arg[A(e^{j\omega T})] = 2\,\arg\,[P(e^{j\omega T})] = -2\,\arg\,[D_{min}(e^{j\omega T})] \tag{II.29}$$

That is, the all-pass system must first cancel the phase of $D_{min}(z)$, which is $-\arg[P(e^{j\omega T})]$, and then add in the phase of $D_{max}(z)$, which is $+\arg[P(e^{j\omega T})]$.

To demonstrate that such an all-pass filter can be designed, we used the window design method [8] to design a finite impulse response (FIR) approximation to the desired all-pass filter. First the frequency response of the ideal all-pass filter was expressed as

$$A(e^{j\omega T}) = \left[e^{j\,\arg[P(e^{j\omega T})]}\right]^{2} \tag{II.30}$$

Note that we do not simply double the phase since if the principal value (PV) of $\arg[P(e^{j\omega T})]$ is computed (as it would be using the ATAN2 subroutine in FORTRAN), then

$$2 \arg[P(e^{j\omega T})] \neq 2\, PV[\arg[P(e^{j\omega T})]] \tag{II.31}$$

unless $-\pi < \arg[P(e^{j\omega T})] < \pi$.

The next step in the design was to sample $A(e^{j\omega T})$ as given by Eq. (II.30) at L equally spaced frequencies giving the sequence

$$A_p(k) = A(e^{j\omega_k T}) \qquad 0 \leq k \leq L-1 \tag{II.32}$$

where

$$\omega_k = \frac{2\pi k}{LT} \tag{II.33}$$

The inverse DFT of $A_p(k)$ was then computed, to obtain

$$a_p(n) = \frac{1}{L} \sum_{k=0}^{L-1} A_p(k) e^{j\frac{2\pi}{L} kn} \tag{II.34}$$

It can be shown [8] that since $A_p(k)$ is a sampled version of $A(e^{j\omega T})$ then

$$a_p(n) = \sum_{r=-\infty}^{\infty} a(n + rL) \tag{II.35}$$

where $a(n)$ is the ideal impulse response of the all-pass filter. That is, $a_p(n)$ is a time-aliased version of the desired impulse response, $a(n)$. If L is large enough, this aliasing is not severe.

As we have pointed out, the ideal all-pass filter for converting the minimum-phase de-emphasis filter into a maximum-phase filter must be non-causal. Thus, a causal FIR approximation requires that $a_p(n)$ be delayed (modulo L) before being truncated by a window function. That is, the causal FIR approximation is

$$\tilde{a}(n) = a_p(n - N_d)w(n). \tag{II.36}$$

Because of the particular properties of the all-pass system, it was found that the best results were obtained with simply a rectangular window.

An example will illustrate the above design procedure. Figure II-14 shows the log magnitude (in dB) and the phase of the minimum-phase de-emphasis system of Eq. (II.26) with r = .8 and θ = .243 radians. The maximum-phase de-emphasis system would, of course, have the same log magnitude function but the phase would be the negative of the function shown in Fig. II-14b. Note that the impulse response of this system is shown in Fig. II-12. Now after sampling the phase function at 1024 points (L = 1024) and forming $A_p(k)$, the aliased ideal impulse response was computed using a 1024-point FFT. It was delayed N_d = 29 samples and truncated to 32 samples to produce the finite impulse response shown in Fig. II-15. To check on the frequency response of the resulting filter, the discrete Fourier transform of the impulse response was computed using a 1024-point FFT. The result is plotted in Fig. II-16, which shows the log magnitude (in dB) and the phase approximation error. It is evident that the magnitude response deviates from 1 (0 dB) by at most about .03 dB and the phase is within .004 radians of the desired $-2\arg[D_{max}(e^{j\omega T})]$.

As further evidence that this is a reasonable approach to maximum phase de-emphasis, Figure II-10f shows the output of the system in Figure II-13b, or equivalently, the result of processing the waveform of Figure II-10c with an all-pass FIR all-pass approximation. Comparing Figures II-10d and II-10f shows very little difference as would be expected from the accuracy of the approximation.

A second approach to designing a maximum-phase de-emphasis system is to use an FIR approximation to the maximum phase de-emphasis filter of

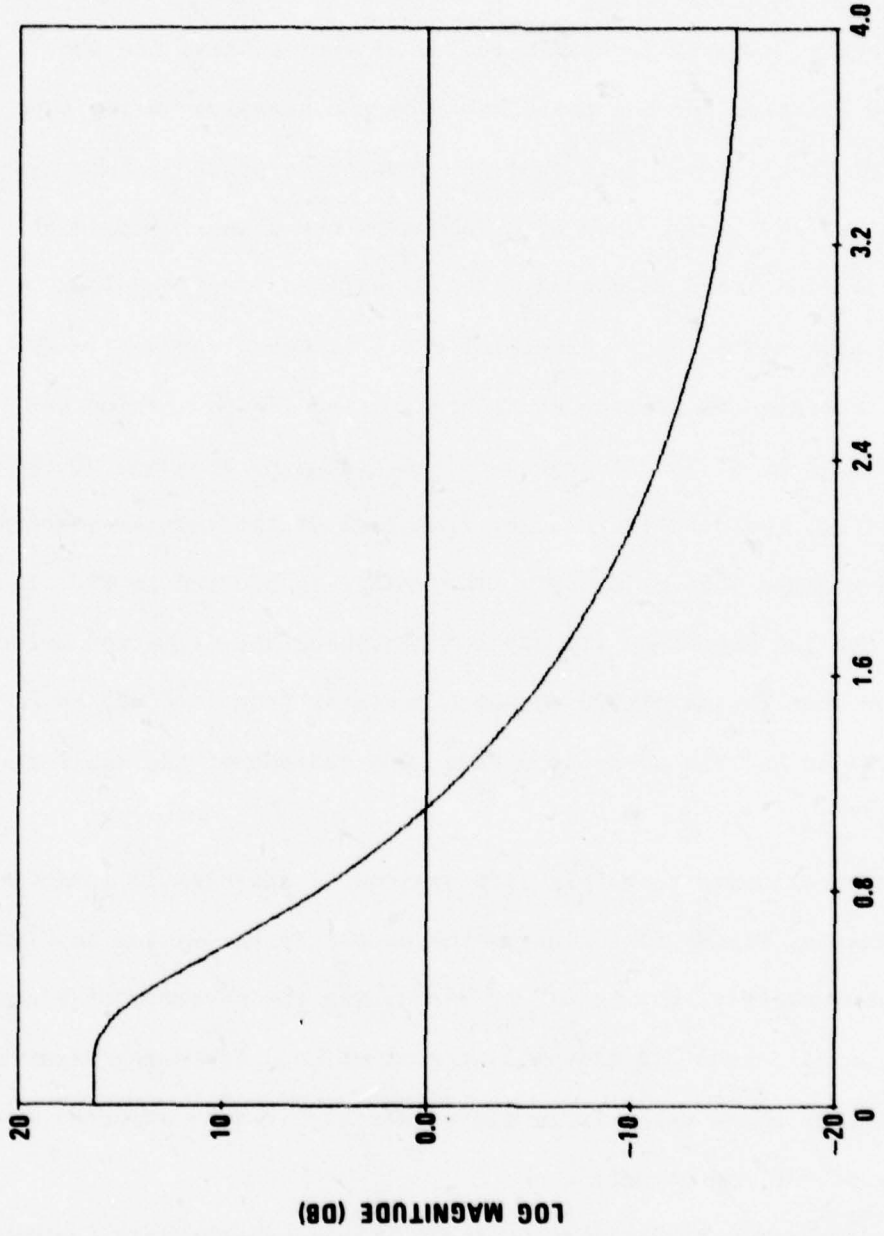


FIGURE II-14a LOG OF MAGNITUDE RESPONSE OF MINIMUM PHASE DE-EMPHASIS FILTER.

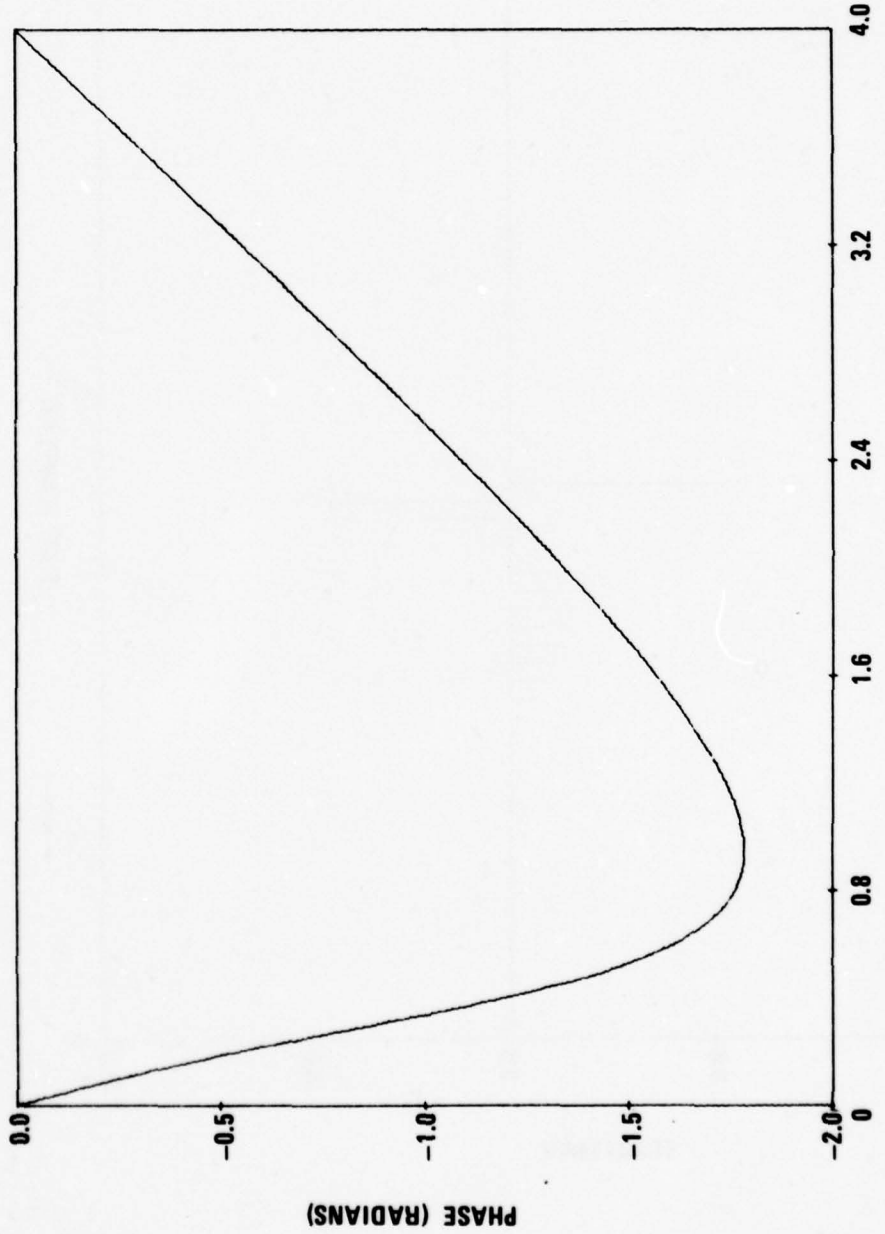


FIGURE II-14b PHASE RESPONSE OF MINIMUM PHASE DE-EMPHASIS FILTER.

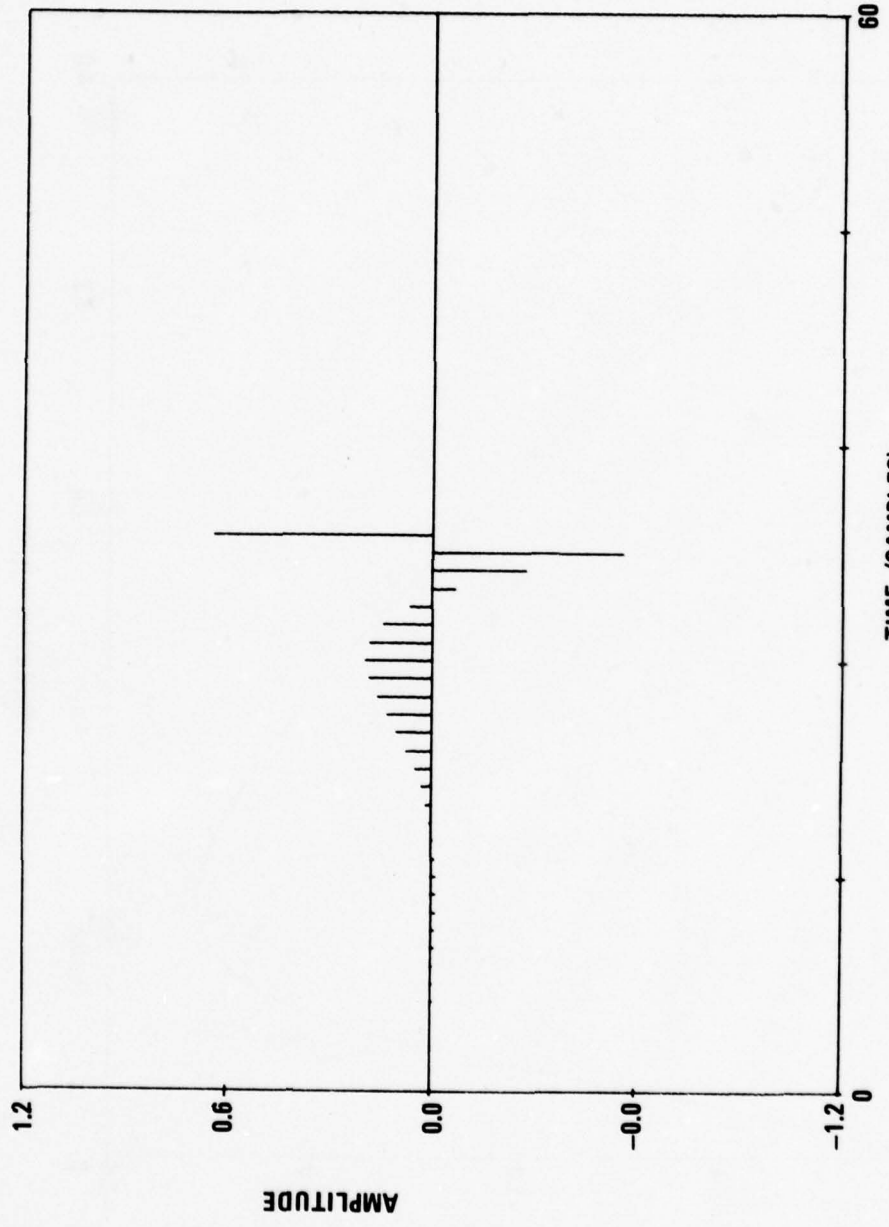


FIGURE II-15 IMPULSE RESPONSE OF FIR ALL-PASS FILTER.

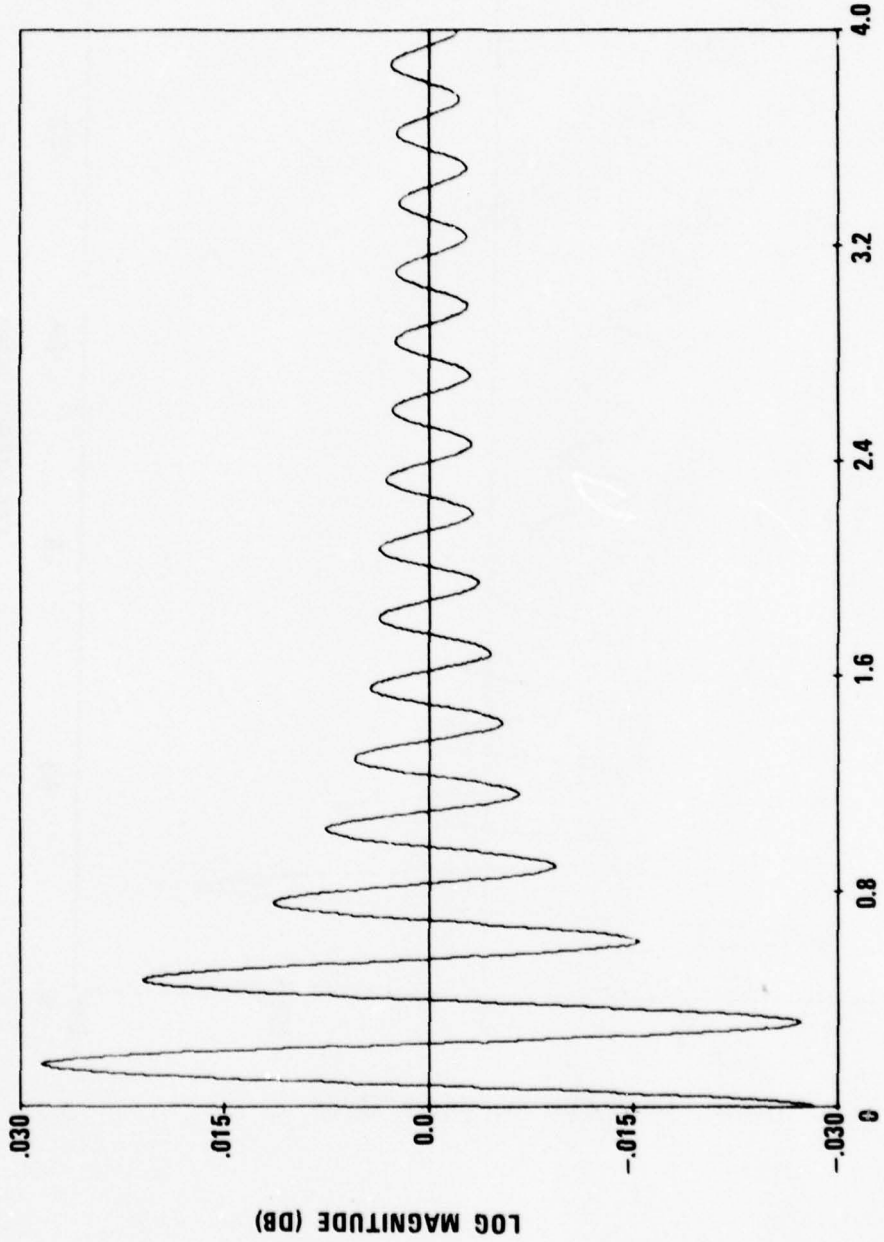


FIGURE II-16a LOG OF MAGNITUDE RESPONSE OF FIR ALL-PASS.

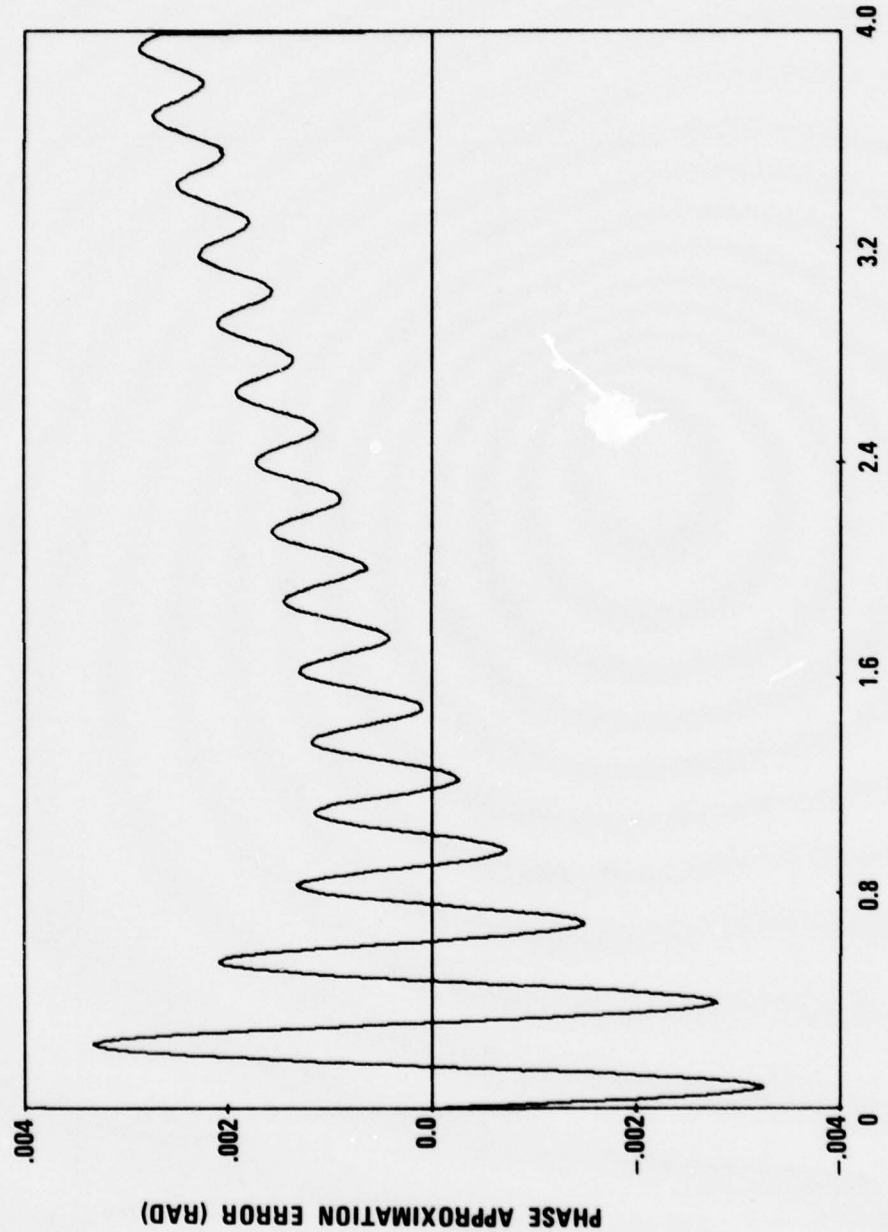


FIGURE II-16b APPROXIMATION ERROR FOR PHASE OF FIR ALL-PASS.

Fig. II-13a. For example, the time-reversed and delayed impulse response of Fig. II-12 could be truncated with a window as in the all-pass design, and the result would serve as an FIR approximation to the maximum phase de-emphasis filter. As another approach, a single "glottal pulse" such as shown in Fig. II-12 was extracted from the inverse filter output for a second order pre-emphasized input. When this pulse was used as the impulse response of an FIR filter in the system of Figure II-13a, the output waveform appeared as in Figure II-10g. Note that the waveshape is very similar to that of the natural speech signal, although somewhat less similar than the waveforms of Figures II-10d-f.

The main draw-back of both the FIR all-pass approximation and the FIR maximum-phase de-emphasis filter is the excessive amount of computation required for their implementation. For this reason, it is desirable to use the lowest order FIR approximations attainable. For the all-pass filter, we found that suitable magnitude and phase response was achieved with an impulse response as short as 32 samples. However, this still requires about 32 multiplies and adds per output sample, making real-time operation difficult if not impossible with the PSP computer.

The specialized structure of the LPC synthesizer permits some approximations which make both the FIR all-pass filter and the FIR maximum-phase de-emphasis filter feasible. We note that for a cascade of linear time-invariant systems, the order of the systems is irrelevant. For time varying systems such as the LPC synthesizer, this is not the case; however, since the time variation in this case is rather slow, it is a reasonable approximation to move the all-pass filter or the de-emphasis filter to a position before the LPC filter. Since the maximum phase de-emphasis is

only necessary for voiced speech, we can move the FIR all-pass in Fig. II-13b to the output of the impulse train generator as shown in Fig. II-17a. Thus, the minimum phase de-emphasis filter, which can be implemented with a simple, second-order, recursive difference equation, remains at the output of the synthesizer whereas the all-pass filter is only applied to the impulse train excitation for voiced speech. Thus, the impulse train generator and the impulse response of the all-pass filter can be combined into a "pulse train generator" thereby eliminating most of the multiplications that would be required to implement the all-pass filter at the output. Figure II-17b shows a similar simplification for an FIR approximation to the de-emphasis filter. In this case, the FIR maximum-phase de-emphasis filter is combined as before with the impulse train generator in the voiced excitation branch. However, in this case the minimum-phase de-emphasis filter must be moved into the unvoiced branch, since both voiced and unvoiced speech was pre-emphasized. (An alternative would be to not pre-emphasize the unvoiced frames.) As in the case of Fig. II-17a, this implementation requires very little extra computation over using only the minimum-phase de-emphasis filter by itself at the output of the synthesizer.

We have thus demonstrated two practical approaches to the implementation of maximum-phase de-emphasis. We now shall consider the benefits of maximum-phase de-emphasis in the context of the LPC-to-CVSD conversion process.

II-4. Signal-to-Noise Ratio Measurements

In order to quantitatively assess the benefits of maximum-phase de-

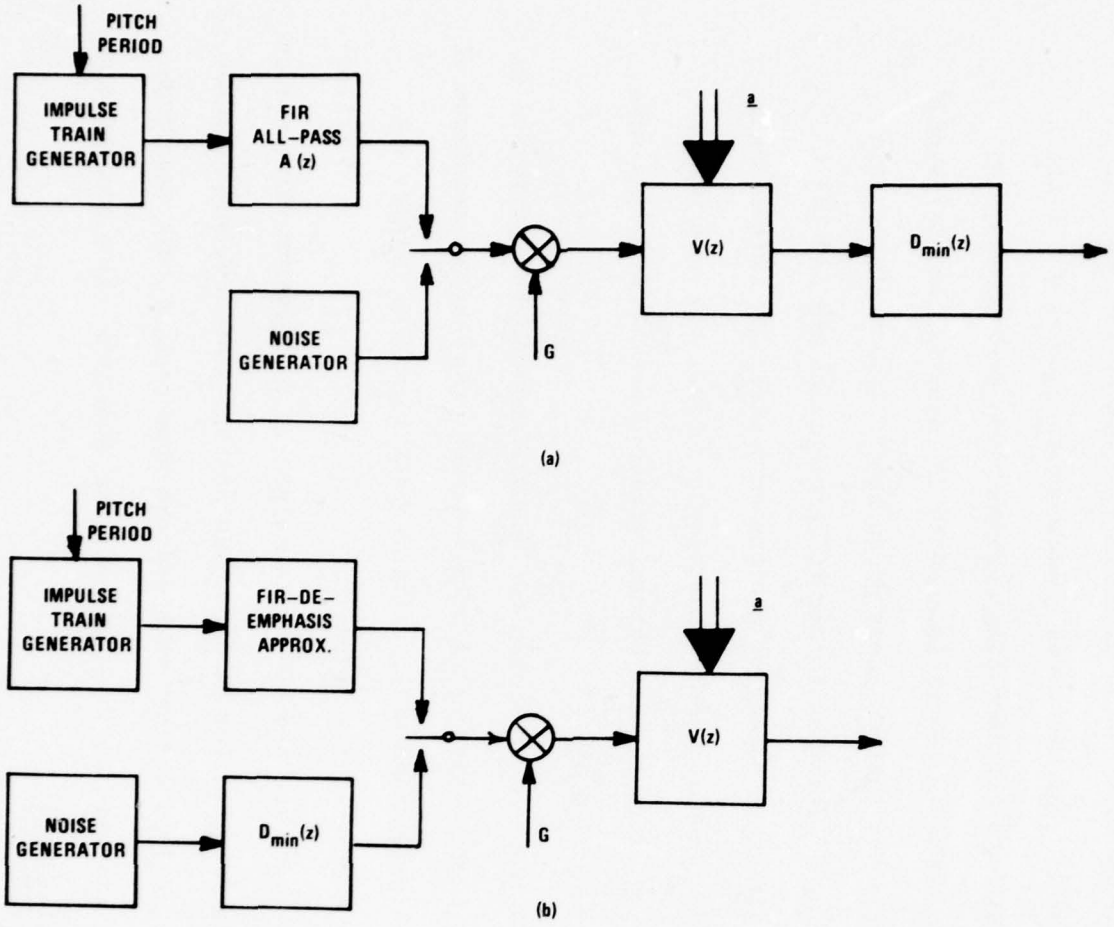


FIGURE II-17 SYNTHESIZER CONFIGURATIONS (a) INCORPORATING FIR ALL-PASS. (b) INCORPORATING FIR MAXIMUM PHASE DE-EMPHASIS.

emphasis in the LPC-to-CVSD conversion process, an extensive set of measurements was made on quantizing noise introduced by the CVSD system for various types of LPC coding as input to the CVSD system. The signal-to-quantizing noise ratio as defined by Figure II-5 and Eq. (II.18), was measured for LPC coded speech obtained using (a) first order pre-emphasis and corresponding minimum-phase de-emphasis, (b) second order pre-emphasis and minimum phase de-emphasis, and (c) second order pre-emphasis and maximum-phase de-emphasis (implemented by time reversal filtering). These measurements were carried out for all 6 sentences at 7 different values of minimum step-size (i.e. effectively 7 different signal levels). For comparison purposes, the same measurements were also obtained for the natural speech input. (The latter data was shown in Figure II-6.)

In order to insure consistency between the measurements performed on the 4 different representations of each sentence, the average magnitude of each version was determined using a first order recursive filter; i.e.

$$M(n) = aM(n - 1) + |x(n)| \qquad \text{(II.37)}$$

where x(n) represents the samples of the signal. The maximum value across the sentence was recorded for each version of that sentence, and then the signals were normalized so that each had the same peak average magnitude. This results in all versions having about the same peak value and sounding about the same loudness.

The complete set of measurements is presented in Figures II-18a - II-18f. It should be stressed from the beginning of this discussion that these curves only give an indication of how well the CVSD part of the tandem connection represents the <u>waveform</u> that is provided as its input. It

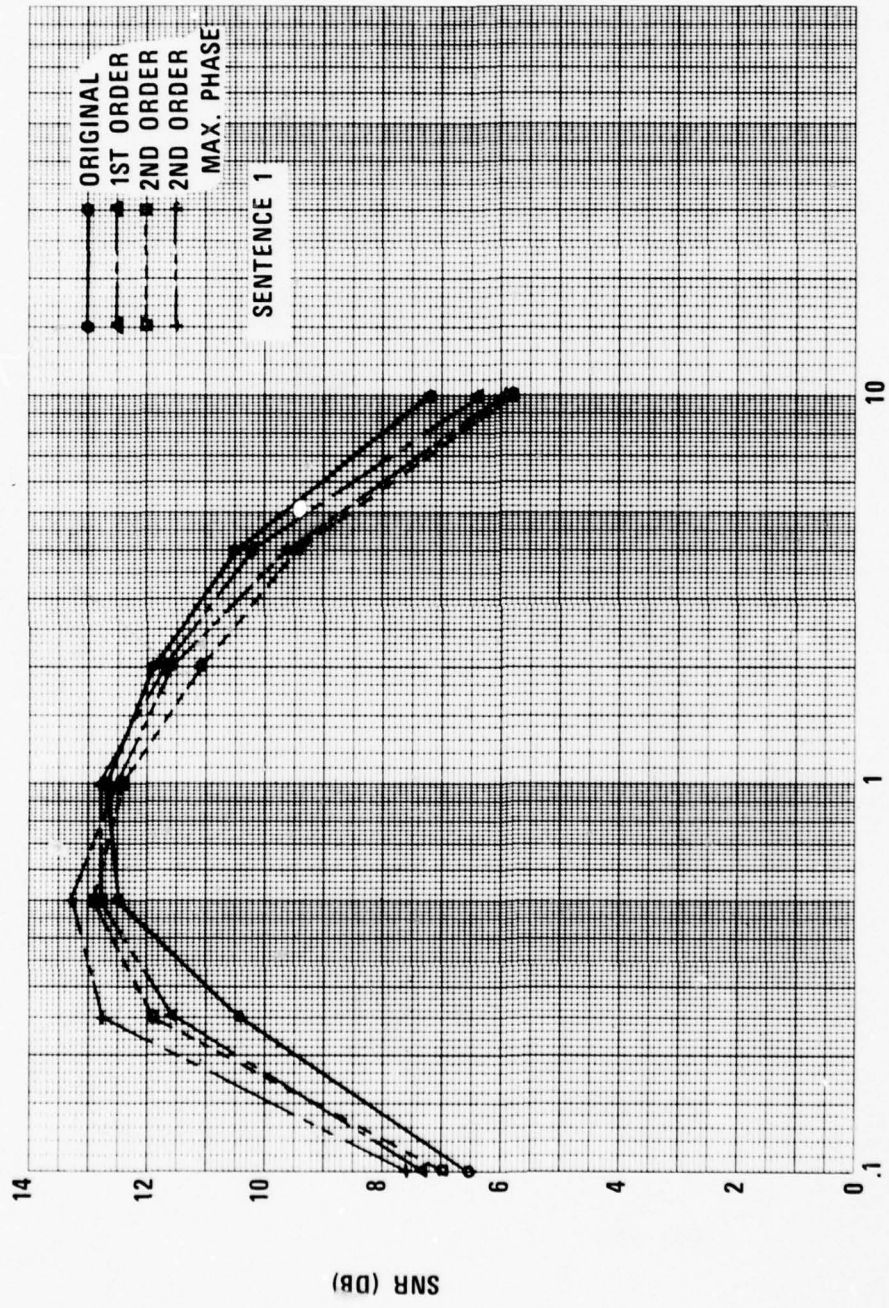


FIGURE II–18a SIGNAL–to–NOISE RATIO AS A FUNCTION OF MINIMUM STEP–SIZE FOR SENTENCE 1.

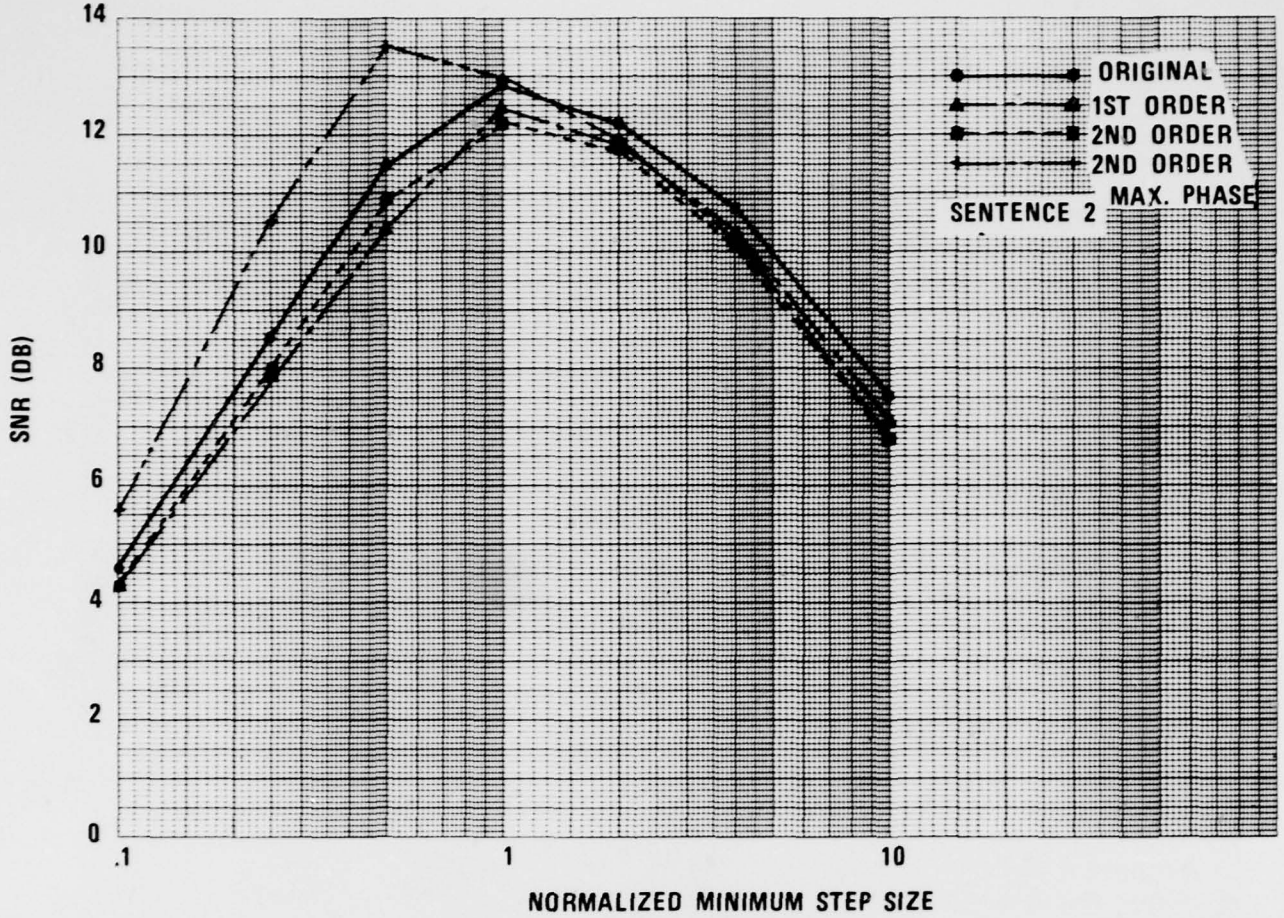


FIGURE II–18b SIGNAL–to–NOISE RATIO AS A FUNCTION OF MINIMUM STEP–SIZE FOR SENTENCE 2.

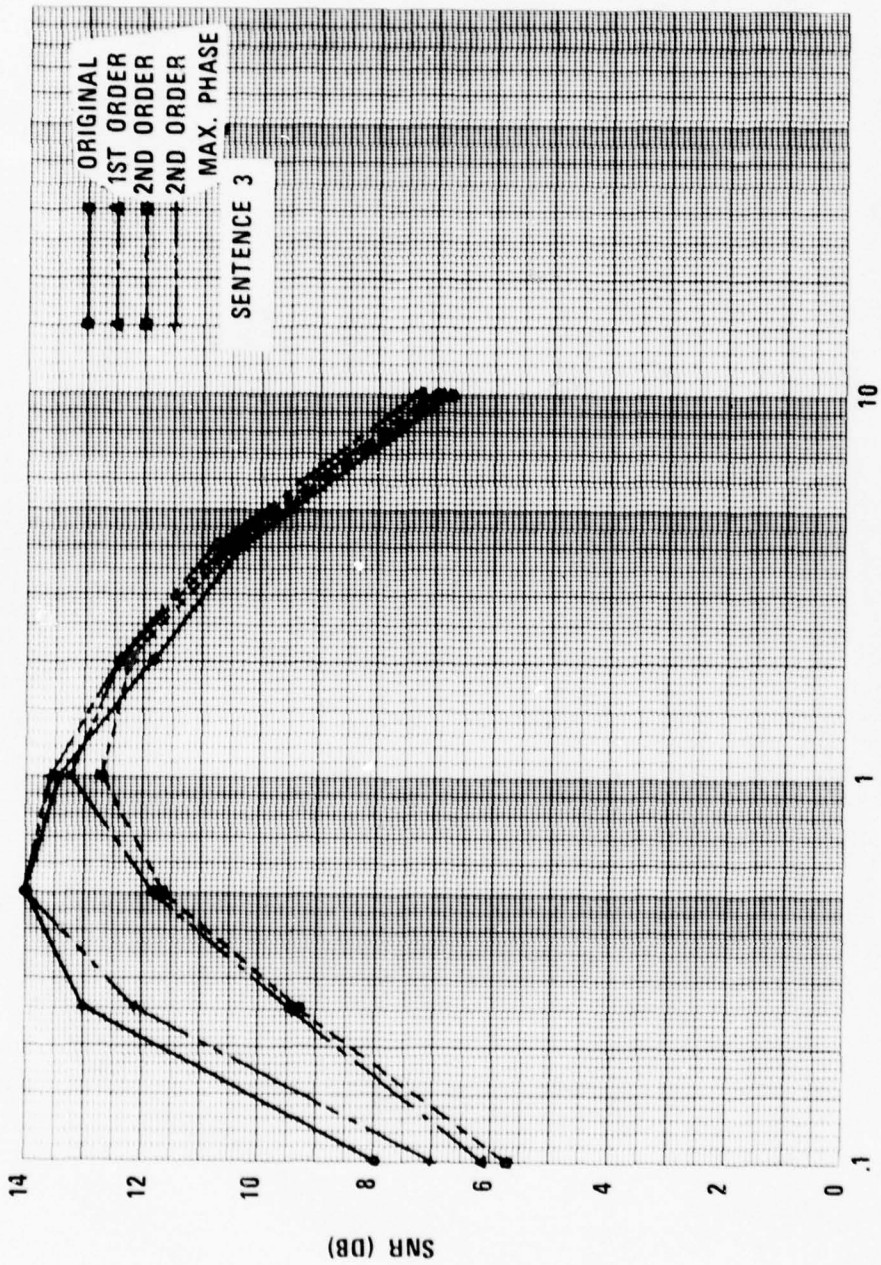


FIGURE II-18c SIGNAL-TO-NOISE RATIO AS A FUNCTION OF MINIMUM STEP-SIZE FOR SENTENCE 3.

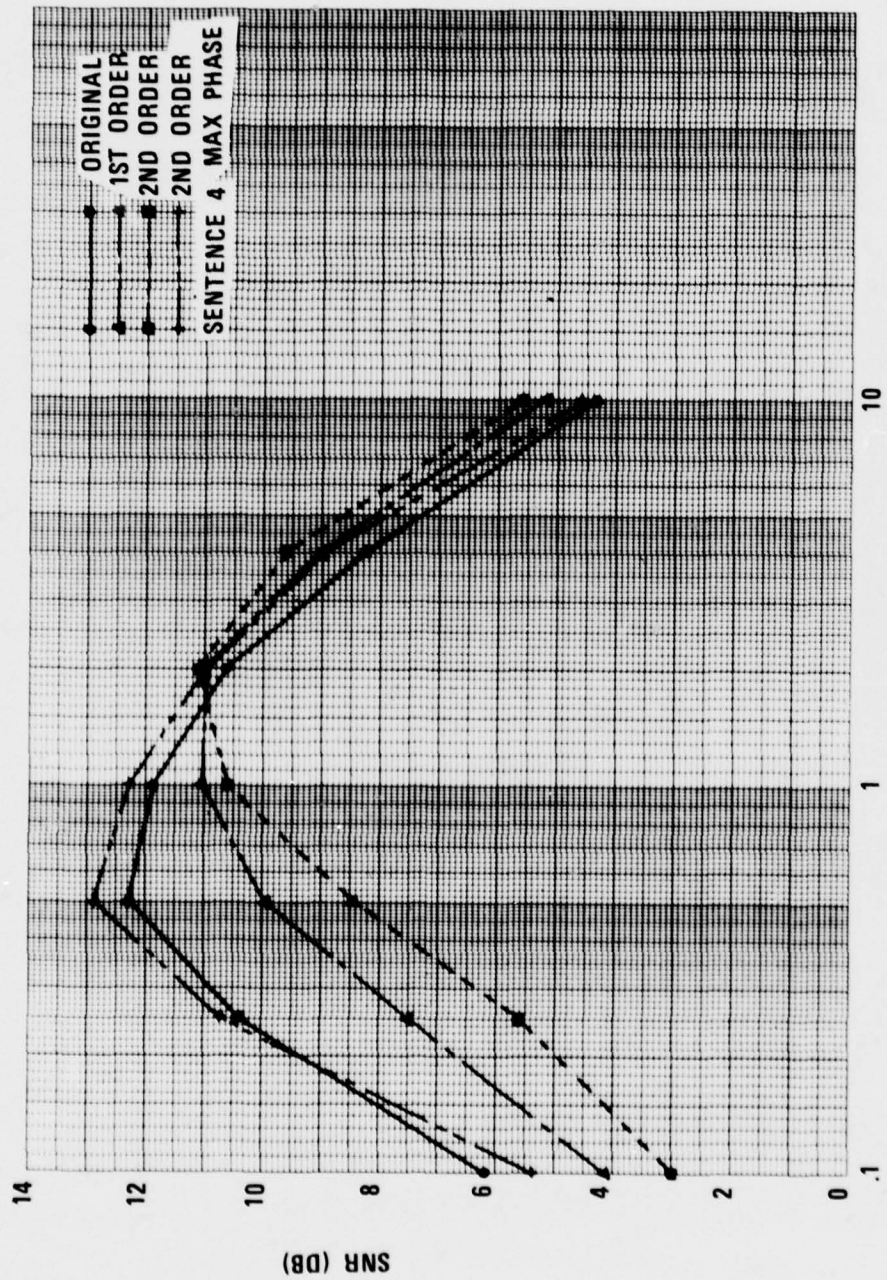


FIGURE II-18d SIGNAL-*to*-NOISE RATIO AS A FUNCTION OF MINIMUM STEP-SIZE FOR SENTENCE 4.

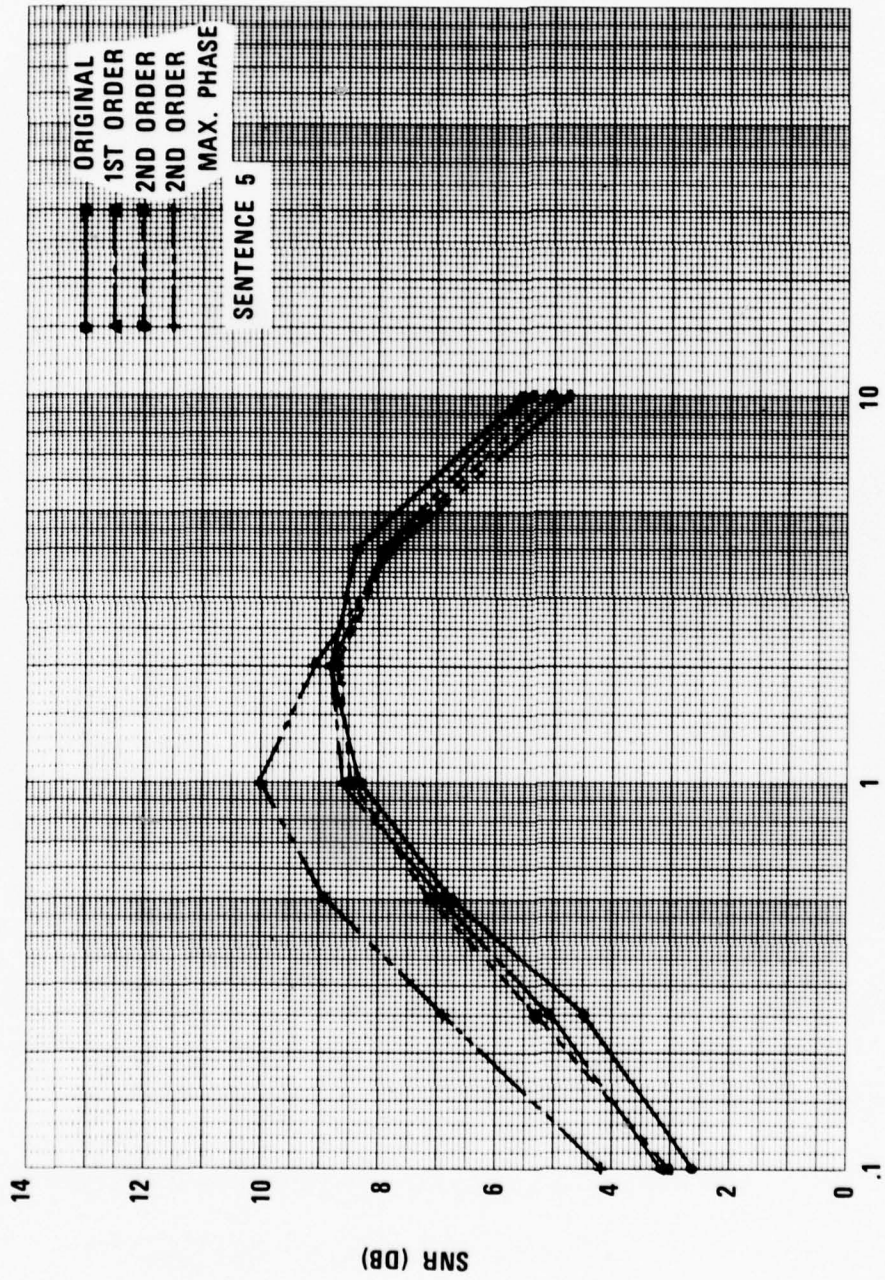


FIGURE II–18e SIGNAL–to–NOISE RATIO AS A FUNCTION OF MINIMUM STEP–SIZE FOR SENTENCE 5.

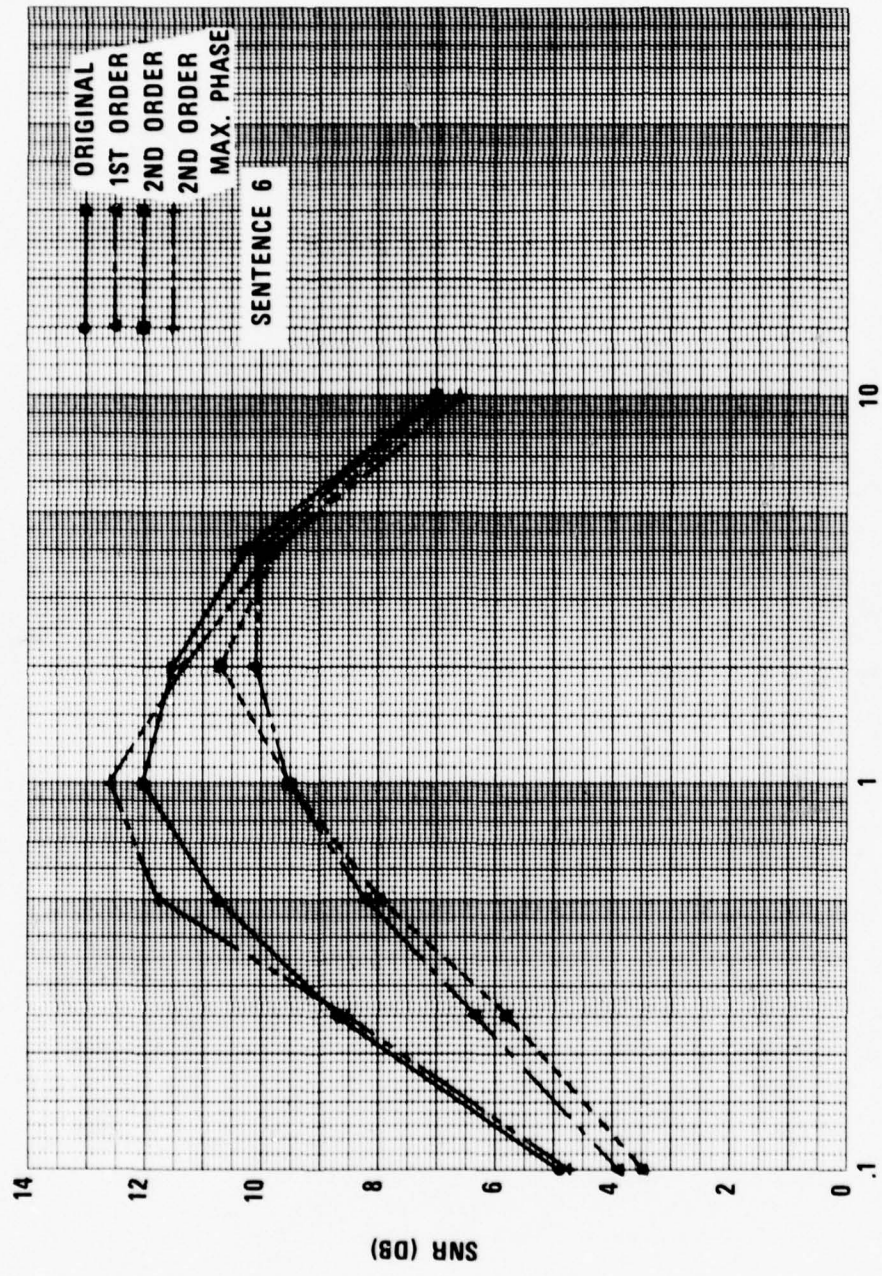


FIGURE II-18f SIGNAL-to-NOISE RATIO AS A FUNCTION OF MINIMUM STEP-SIZE FOR SENTENCE 6.

is, of course, meaningless to consider the signal-to-noise ratio of the overall LPC-to-CVSD connection since the LPC system is not designed to preserve waveform shape.

Careful scrutiny of these graphs leads to the following conclusions.

(1) The general behavior of CVSD is the same regardless of whether the input is natural speech or LPC coded speech. That is, the SNR is a maximum for normalized step sizes in the range .5 to 2 and it falls off for small step sizes (slope overload) and for large step sizes (granular noise).

(2) The SNR for all forms of LPC-coded speech is sometimes better than for natural speech. This obviously does not mean that the output in these cases is of higher subjective quality when compared to the original utterance.

(3) The SNR curves come together for large step sizes. This is consistent across all 6 sentences/speakers. The main differences in the curves occur for small step sizes (i.e. in the slope overload region).

(4) In the slope overload region, the SNR for LPC coded speech with second order pre-emphasis and maximum phase de-emphasis is consistently higher than with either of the other forms of LPC coding. This is true for all six sentences.

The data displayed in Figures II-18a - II-18f indicate that the second order maximum-phase de-emphasis has a significant effect on the performance of the CVSD system. Since this form of LPC coding is perceived to be at least as good as the other forms investigated, it is reasonable

to suppose that the output of the CVSD system should sound at least as good in the maximum-phase case as in the other cases. However, it is well known that SNR is not a good measure of subjective quality in delta modulators. Thus, a perceptual test was performed to assess subjective quality differences.

II-5. Perceptual Evaluation

To determine whether any of the system configurations studied produce a perceptually significant improvement in quality, a formal subjective listening test was performed. Since, at this time, the subjective testing procedure which has been most throughly tested is the PARM (Paired Acceptability Rating Method) test [2], a procedure very similar to PARM was used. The complete details of the test format are given as Appendix C.

An innovation in the subjective testing was that the digital signal processing facility was used not only to prepare the material for the test, but to administer the test, collect the data, and run the statistics on the results. The details of the speech quality testing facility are given in Appendix D.

II-5.1. The Design of the Perceptual Tests

The purpose of the perceptual tests was to quantify the improvements, if any, caused by the phase modification techniques, discussed above. The format of the test was similar to a PARM test([2], Appendix C), with some differences. As in the PARM, subjects were asked to listen to 60 sentences processed by six systems: a high anchor, a low anchor, and four test systems. However, in this test, only five sentences, each

spoken by a different speaker, were used. (The first five sentences of Appendix A.) Each of the sentences was processed by each of the systems, for a total of 5x6=30 sentences. As in the PARM, the systems were paired in all possible ways, and both the forward pairing and the backward pairings were used. However, in these tests, an additional constraint was added so that all possible pairings of sentences in both directions were also used. Hence, during the test, the subjects heard all possible system combinations and all possible sentence combinations paired in both directions. Another minor variation from the PARM test was that the sentences were explicitly presented in pairs. Subjects were asked to listen to two sentences, and then to key in their two digit responses for each. As in the PARM, subjects were instructed to key in answers between 0 and 99 with five points resolution.

In all, four subjective tests were designed for studying the LPC-to-CVSD tandem. Each test was performed at a particular minimum step size, and each contained four systems: CVSD alone; CVSD of LPC with first order pre-emphasis and minimum-phase de-emphasis; CVSD of LPC with second order pre-emphasis and minimum-phase de-emphasis, and CVSD of LPC with second order pre-emphasis and maximum-phase de-emphasis. In all cases, the LPC simulations and CVSD simulations were the same as those described earlier in this part.

In each case, eighteen subjects were given the subjective test. No subject related corrections in the data were performed, and the only corrections of any type which were performed were data screenings for keying mistakes. The statistical analysis used was the Newman-Keul test,

which is described in Appendix C.

A last important point should be made concerning these tests and those described in Section III-6. The "low anchor" used in this study was of considerably better quality than that used in the original PARM tests [2]. This resulted in the overall scores being biased down in this test. Hence, great care should be exercised when comparing these test results to previous PARM results.

II-5.2 The Subjective Quality Results

Table II-2 shows a compilation of the results for the four tests, while Table II-3 shows the results of a statistical analysis of the data in Table II-2. In the statistical study, the systems were first ranked by means, and the differences in means were tabulated. These differences were then presented in matrix format above the diagonal as shown in Table II-3. Below the diagonal, a blank means the corresponding difference in means was not significant at either the .01 or .05 level, a "*" means the difference is significant at the .05 level, while a "**" means the difference is significant at the .01 level.

Several points should be made concerning these results. First, as reported by others [3], there is a marked preference among the subjects for the slope-overload condition, and a marked rejection of the granular case (normalized minimum step size = 4). Second, of the groups tested, the best overall result was obtained by a minimum step size of .5. Third, for the two end cases (minimum step size = .1 and 4), the phase modification makes very little difference in the perceived results. Fourth, in the middle two cases the maximum-phase modification (System 5) gives a statistically significant improvement over both the minimum-phase cases

(Systems 3 and 4). The improvements range from 2.0 points to 4.8 points, and, at a minimum step size of .5, the improvement is enought so that the maximum phase de-emphasis case is not found to be significantly different from CVSD alone.

In summary, therefore, it can be said that the results here mirror very closely the results of the signal-to-noise ratio study. At the extremities (very strong talkers and very weak talkers), these techniques have little effect. In the center, however, the improvements exist in a statistically significant sense, but these improvements are not large.

		Normalized Minimum Step-Size			
		.1	.5	2	4
1.	High Anchor	73.3	74.8	73.8	74.4
2.	CVSD	47.9	48.6	45.7	39.4
3.	LPC + CVSD (1st order pre-emphasis minimum phase de-emphasis)	41.8	42.4	40.8	36.1
4.	LPC + CVSD (2nd order pre-emphasis minimum phase de-emphasis)	40.8	42.2	39.2	34.8
5.	LPC + CVSD (2nd order pre-emphasis, maximum phase de-emphasis)	41.1	44.4	44.0	34.8
6.	Low Anchor	25.3	24.2	25.9	24.1

Table II-2 Means for the subjective tests for the LPC-to-CVSD tandem.

Normalized Minimum Step-Size

.1

	HI	2	3	5	4	Low
HI		25.4	31.5	32.2	32.5	48.0
2	**		6.1	6.8	7.1	22.6
3	**	**		.7	1.0	16.5
5	**	**			.3	15.8
4	**	**				15.5
Low	**	**	**	**	**	

.5

	HI	2	5	3	4	Low
HI		26.2	30.5	32.5	32.7	50.6
2	**		4.3	6.3	6.5	24.4
5	**	**		2.0	2.2	21.0
3	**	**	*		.2	18.1
4	**	**	**			18.0
Low	**	**	**	**	**	

2

	HI	2	5	3	4	Low
HI		28.1	29.8	33.0	34.6	47.9
2	**		1.7	4.9	6.5	19.8
5	**			3.3	4.8	18.1
3	**	**	**		1.6	14.9
4	**	**	**			13.3
Low	**	**	**	**	**	

4

	HI	2	3	5	4	Low
HI		35.1	38.3	39.6	39.6	50.4
2	**		3.3	4.6	4.6	15.3
3	**	**		1.3	1.3	12.0
5	**	**			0	10.7
4	**	**				10.7
Low	**	**	**	**	**	

Table II-3. Results of the Newman-Keul test on the four subjective quality tests.
"*" Means significance at the .05 level.
"**" Means significance at the .01 level.

PART III

CVSD-to-LPC Tandem Connection

III-1. Simulation of LPC-to-CVSD Connection

The components of the LPC-to-CVSD tandem connection as shown in Figure I-1b. include: (a) the CVSD coder, which generates a 16 Kbit/sec digital representation of the speech signal, (b) a system for converting from the CVSD representation to a low bit-rate LPC representation, and (c) a synthesizer for converting the LPC representation to an analog waveform for listening. The components of the simulation of this system are shown in Figure III-1. It is clear from a comparison of Figures II-1 and III-1, that essentially the same components are required in the simulation of both tandem connections. These components are simply cascaded in different order in simulating the two directions of conversion. Indeed, the discussion of Section II-1 is sufficient to define the nature of the simulation of the CVSD-to-LPC tandem connection of Figure III-1, except for modifications to the LPC analysis algorithm which will be described in detail in Section III-4.

It can be seen that the simulation of the CVSD-to-LPC conversion process is composed of a CVSD decoder and an LPC analyzer. That is, it is assumed that a waveform representation must first be obtained from the CVSD bit-stream before the LPC representation is computed.* The waveform so obtained will, of course, be contaminated with CVSD quantization noise. The effect of this quantization noise on the computation of the LPC representation is a major factor in the performance of the CVSD-to-LPC tandem connection.

---

*It is possible that pitch could be adequately estimated directly from b(n), but unlikely that the LPC coefficients could be estimated without first obtaining the quantized waveform $\tilde{x}(n)$

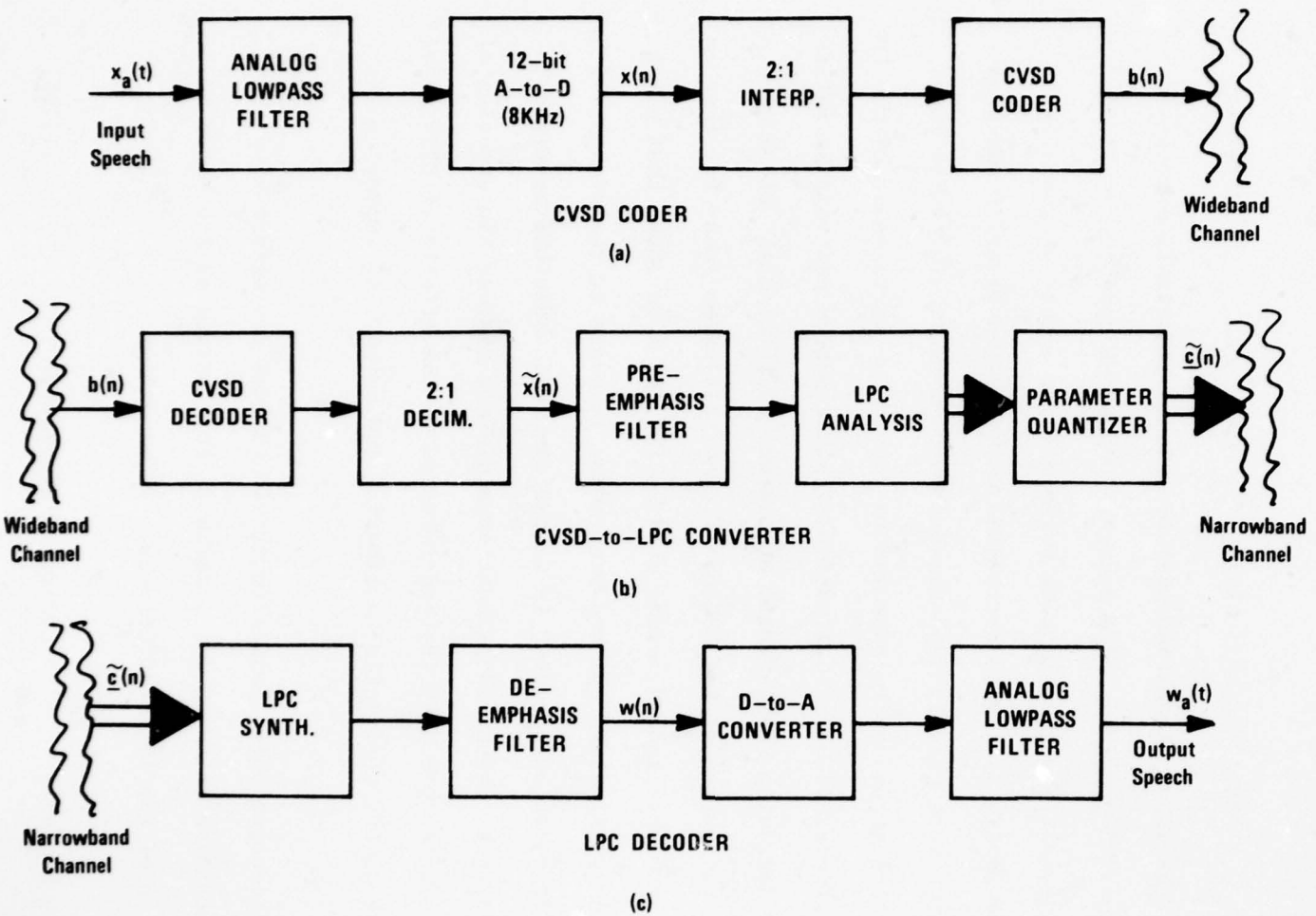


FIGURE III–1 BLOCK DIAGRAM REPRESENTATION OF CVSD–to–LPC TANDEM CONNECTION SIMULATION. (a) CVSD CODER (b) CVSD–to–LPC CONVERTER (c) LPC DECODER.

III-2. Performance of the CVSD-to-LPC Tandem Connection

The signal obtained by decoding the CVSD bit-stream can be represented as

$$\tilde{x}(n) = x(n) + e(n) \qquad \text{(III.1)}$$

where x(n) is the input signal and e(n) represents quantization noise introduced by the CVSD coder/decoder system. Depending upon the amplitude of the input signal, the quantization noise will be either of the slope overload type (which is highly correlated with the input) or of the granular type (which more closely approximates white noise). In either case, this quantization noise can be expected to impair the performance of the LPC coder by interfering with the estimation of pitch period and voicing information and by degrading the estimate of the LPC coefficients. Of these effects, the degradation of the LPC coefficient estimates is by far the more serious problem. Indeed, other studies [15,16] have shown that pitch detection is only slightly impaired when noise is added to the signal, with the greatest degradation occuring in voiced/unvoiced decisions as might be expected. For this reason, our simulations did not address the question of pitch and voicing errors, but instead focussed on the effects of CVSD quantization noise on LPC coefficient estimation and on ways of improving the LPC estimate.

The quantization noise manifests itself through its effect upon the short-time autocorrelation estimate which in turn serves as the basis for the computation of the LPC coefficients. If we define

$$R_{\tilde{x}\tilde{x}}(m) = \sum_{n=0}^{L-|m|-1} \tilde{x}(n)w(n)\tilde{x}(n+m)w(n+m) \qquad \text{(III.2)}$$

to be the short-time autocorrelation function of the noisy signal $\tilde{x}(n)$, then by substituting Eq. (III.1) into (III.2) it can easily be shown that

$$R_{\tilde{x}\tilde{x}}(m) = R_{xx}(m) + R_{ex}(m) + R_{ex}(-m) + R_{ee}(m) \qquad \text{(III.3)}$$

where

$$R_{xx}(m) = \sum_{n=0}^{L-|m|-1} x(n)w(n)x(n+m)w(n+m) \qquad \text{(III.4)}$$

and

$$R_{ee}(m) = \sum_{n=0}^{L-|m|-1} e(n)w(n)e(n+m)w(n+m) \qquad \text{(III.5)}$$

are the short-time autocorrelation functions of the signal and quantization noise respectively, and

$$R_{ex}(m) = \sum_{n=0}^{L-m-1} e(n)w(n)x(n+m)w(n+m) \qquad \text{(III.6)}$$

is the short-time cross correlation function between the quantization noise and the signal.

From the above equations it can be seen that the signal and noise interact in a rather complicated way in the short-time autocorrelation function. The effect of this on the LPC estimate is illustrated by Figures III-2 through III-5, each of which shows vocal tract transfer function estimates

$$|V(e^{j2\pi fT}) = \frac{1}{\left| 1 + \sum_{k=1}^{N} a_k e^{-j2\pi fTk} \right|} \qquad \text{(III.7)}$$

for the original speech utterance and several noise conditions. Figure II-2 is for a segment of voiced speech with additive white noise at several signal-to-noise ratios. Figure III-3 is for a segment of unvoiced speech for the

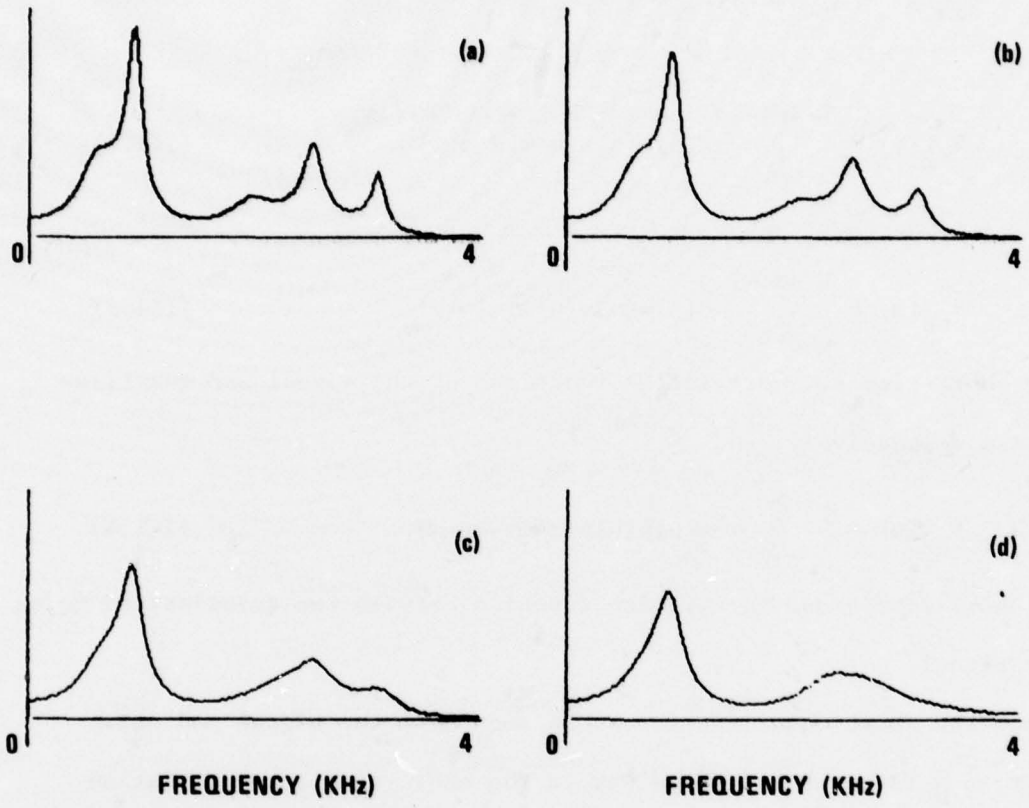


FIGURE III–2 LPC MAGNITUDE SPECTRA (VOICED) (a) NO NOISE (b) ADDITIVE WHITE NOISE, SNR = 20.5. (c) SNR = 14.5 (d) SNR = 8.5

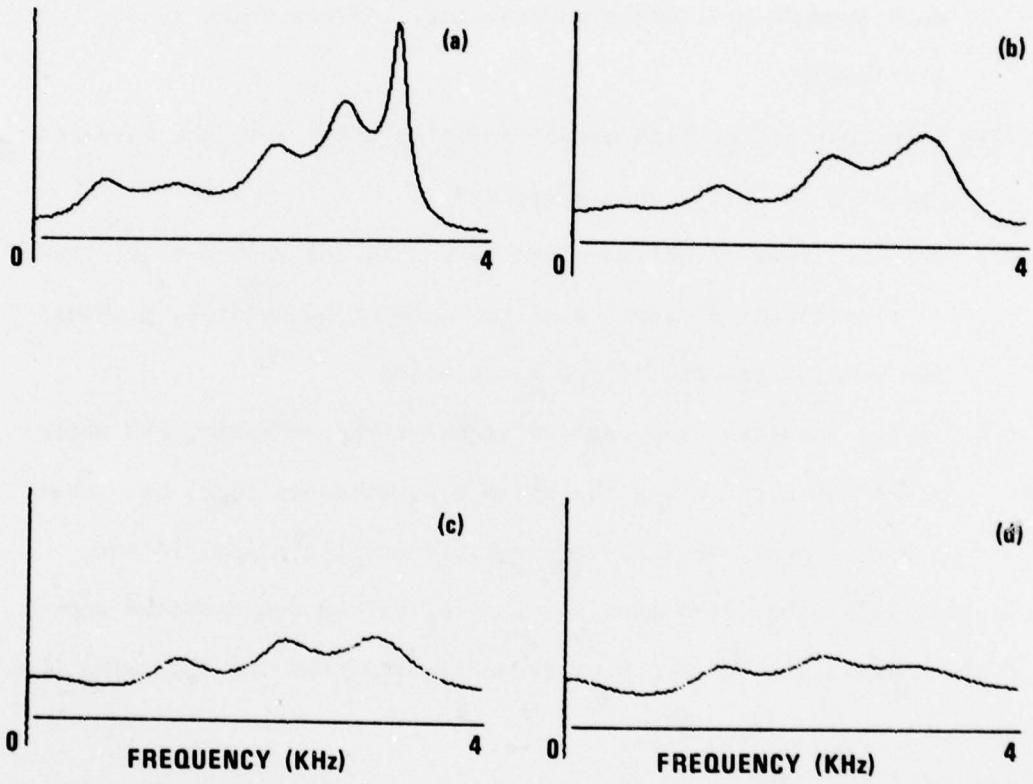


FIGURE III–3 LPC MAGNITUDE SPECTRA (UNVOICED) (a) NO NOISE (b) ADDITIVE WHITE NOISE, SNR = 20.5 (c) SNR = 14.5 (d) SNR = 8.5

same noise conditions. Although our interest is in CVSD quantization noise, these examples for white noise provide an interesting reference point. From Figures III-2 and III-3, the following observations can be made:

(1) First note that the signal-to-noise ratios specified for each example are long-time averages, and the noise is stationary.

(2) Note that at the high signal-to-noise ratio even the unvoiced spectrum is fairly well preserved.

(3) For the lower signal-to-noise ratios in the voiced case, there is significant broadening of the formant bandwidths, although the general spectral shape is retained.

(4) In the unvoiced case, at low signal-to-noise ratio, the white noise dominates since the noise is stationary (i.e. has about the same peak amplitude throughout) and the signal is not.

Figures III-4 and III-5 show results for voiced and unvoiced speech with CVSD quantization noise. From these figures come the following observations:

(1) For the voiced speech case, with the small minimum step-size (slope overload condition) the higher formants are completely obliterated in the LPC spectrum.

(2) For voiced speech with a minimum step-size that gives about the best signal-to-noise ratio, the bandwidths of all the formants are very much broadened. Also note the noise that has entered at high frequencies.

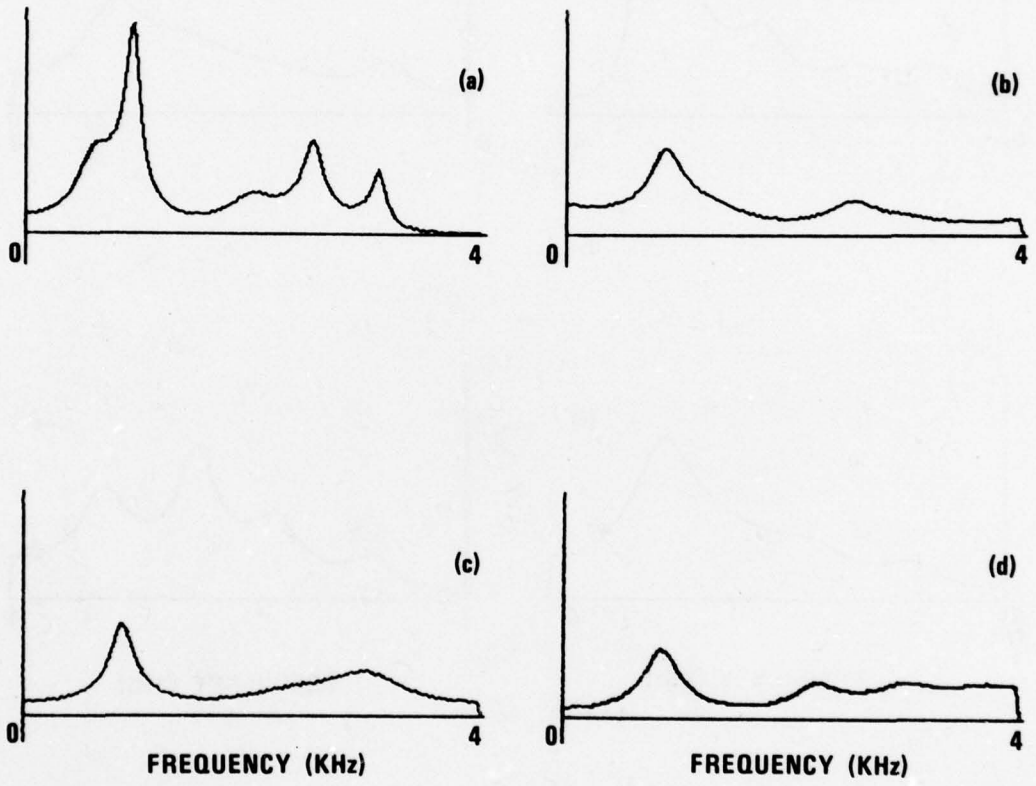


FIGURE III–4 LPC MAGNITUDE SPECTRA (VOICED) (a) NO NOISE (b) CVSD NOISE, $\Delta_{min} = 0.1$ (c) $\Delta_{min} = 1.5$ (d) $\Delta_{min} = 4.0$

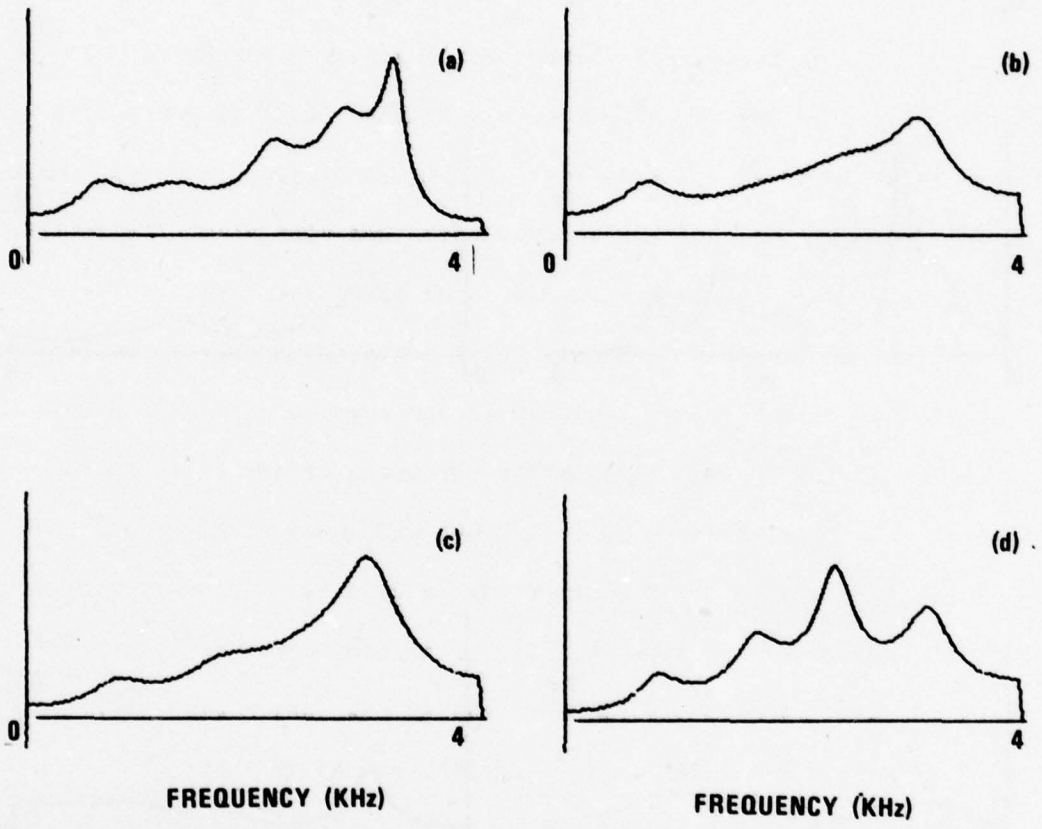


FIGURE III–5 LPC MAGNITUDE SPECTRA (UNVOICED) (a) NO NOISE (b) CVSD NOISE $\Delta_{min} = 0.1$ (c) $\Delta_{min} = 1.5$ (d) $\Delta_{min} = 4.0$

(3) For voiced speech with a large minimum step-size (granular condition), some vestiges of the higher formants remain but there is considerable noise at all frequencies.

(4) In the case of unvoiced speech, it is clear that the effect is much different from the stationary white noise case. Since the noise level tends to follow the speech level, the quantization noise does not overcome the signal and the general spectral shape is reasonably well preserved, especially in the mid-range of minimum step sizes. Although the spectral peaks are broadened as in the voiced case, this effect is likely to be of less importance for unvoiced speech.

The spectral effects illustrated by Figures III-4 and III-5 correspond to definite perceptual degradations in the LPC coded speech. As before, it is helpful to stress the trivial, yet crucial, point that if the CVSD coder could represent the input speech with high accuracy, then the tandem connection would sound as good as LPC coding alone, and we cannot expect to obtain better performance than this without drastic changes in both systems. Since it is unlikely that a better waveform representation can be obtained at 16 kbits/sec using the basic CVSD algorithm as defined above, our efforts were focussed on means of overcoming the effects of the CVSD quantization noise upon the estimate of the LPC coefficients. Two approaches were explored in some depth as discussed below.

III-3. Investigation of Corrections to the Autocorrelation Function

If the terms $R_{ex}(m) + R_{ex}(-m) + R_{ee}(m)$ could be removed in Eq. (III.3), the LPC coefficients could be computed without degradation. This, or course, requires knowledge of the time-varying statistical properties of the noise. Indeed, we must know both the short-time autocorrelation of the noise and

the short-time cross correlation between the noise and the signal. In analyzing the behavior of differential quantizers (of which CVSD is a simple example) it is common to assume that the quantization noise is independent of the signal, and if the quantization is sufficiently fine, it is often further assumed that the quantization noise is white. Under these conditions it would seem that a correction to the autocorrelation function might be possible since the cross correlation error terms might be neglected leaving only the error autocorrelation to estimate. These conditions are sometimes valid for stationary signals but not for CVSD quantization of speech where the properties of the signal and noise change with time and where the noise may be highly correlated with the signal. Nevertheless, in order to gauge the difficulties faced in obtaining an improved LPC estimate from CVSD coded speech, an experiment was carried out to investigate the effects of the various terms in Eq. (III.3).

The different components of Eq. (III.3) were measured as depicted in Figure III-6. (Note that we are concerned with the short-time autocorrelation function so that the operations of Fig. III-6 are carried out on each frame.) Three examples are shown in Figures III-7. For each case, this figure shows the autocorrelation function for the signal with and without noise as well as the autocorrelation function of the noise and the cross correlation terms. It can be seen from these figures that in some cases (e.g. case 1) the cross correlation terms are quite small compared to the signal and noise correlation terms. However, in other cases (e.g. case 2) the cross correlation terms are comparable in size to the autocorrelation terms. That this should occur is not surprising even if

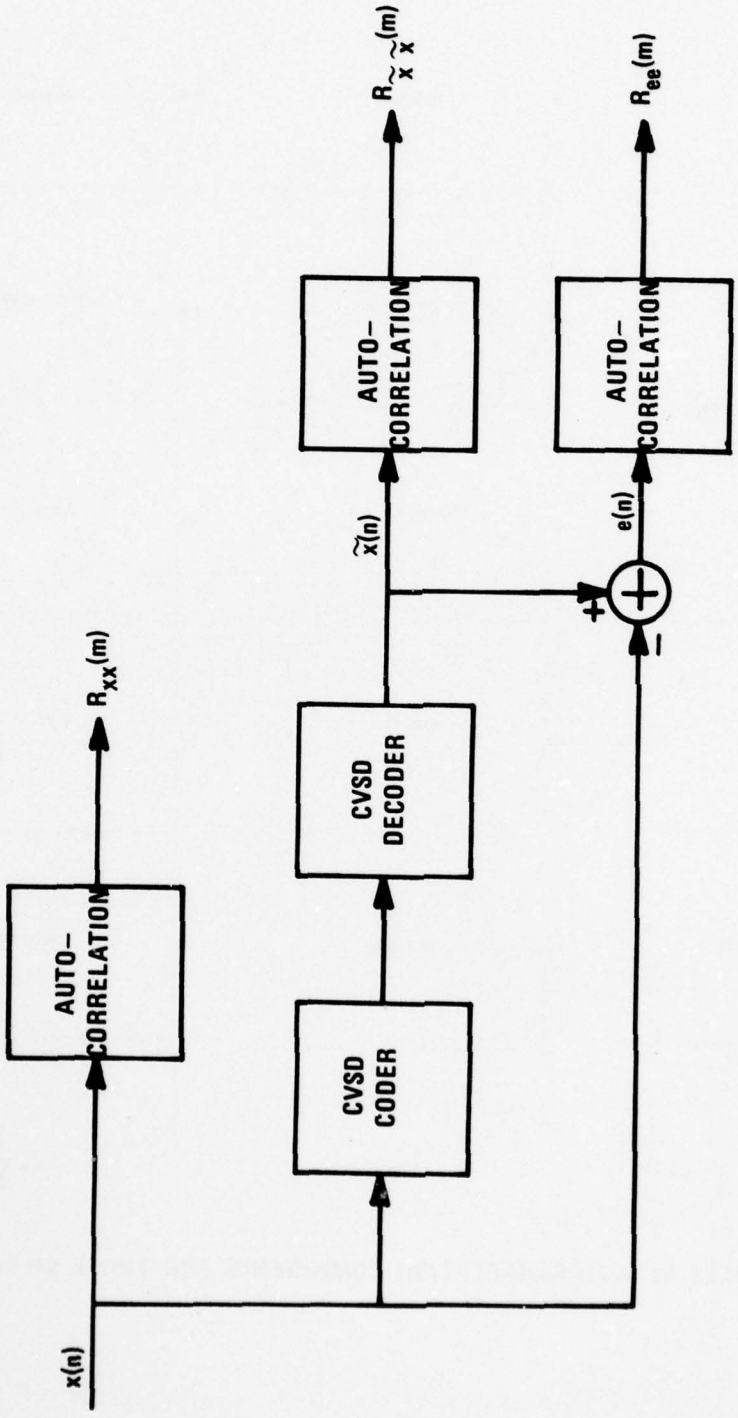


FIGURE III-6 BLOCK DIAGRAM REPRESENTATION OF AUTOCORRELATION FUNCTION MEASUREMENTS.

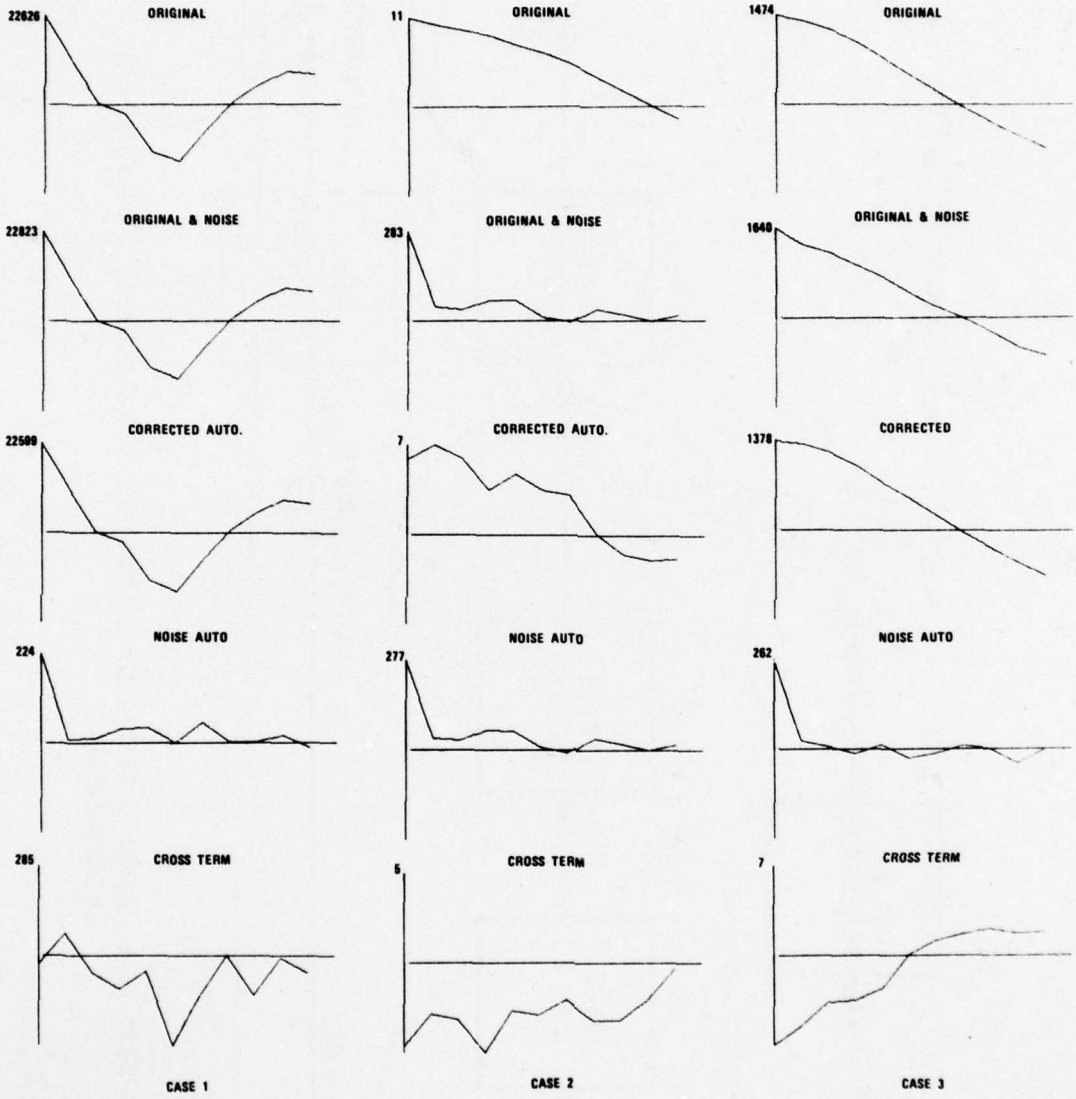


FIGURE III–7 EXAMPLES OF AUTOCORRELATION COMPONENTS FOR THREE SPEECH SEGMENTS.

the signal and noise are completely independent, since the short-time cross correlation function is computed using finite segments of the waveform. Indeed it would be remarkable if the short-time cross correlation terms did turn out to be zero.

The results of Figure III-7 suggest that direct correction of the autocorrelation function is probably not feasible, even if the noise statistics are known. Evidence of this assertion is provided by Figures III-8, III-9, and III-10, which correspond to cases 1, 2, and 3 of Figure III-7, respectively. In each case, the measured autocorrelation function of the noise was subtracted from the autocorrelation function of the CVSD signal. The result was used as the input to the LPC coefficient calculation. The resulting spectra and z-plane pole locations are shown in Figures III-8, III-9, and III-10. It can be seen that in case 1, where the cross correlation terms were relatively small, the "correction" was successful in improving the estimate of the LPC coefficients. However, in the other two cases, poles were forced outside the unit circle by this approach. In observing such results across several sentences, it was found that unstable results occur very frequently, although improvements were also often noted.

Thus, even when the autocorrelation function of the noise is known, it appears that corrections of the autocorrelation function prior to LPC analysis are impractical. A fixed average correction is certain to be even less satisfactory. The behavior displayed in Figure III-8 - III-10 can be attributed to the fact that the autocorrelation function has certain well known special properties which are satisfied by $R_{xx}(m) + R_{ex}(m) + R_{ex}(-m) + R_{ee}(m)$

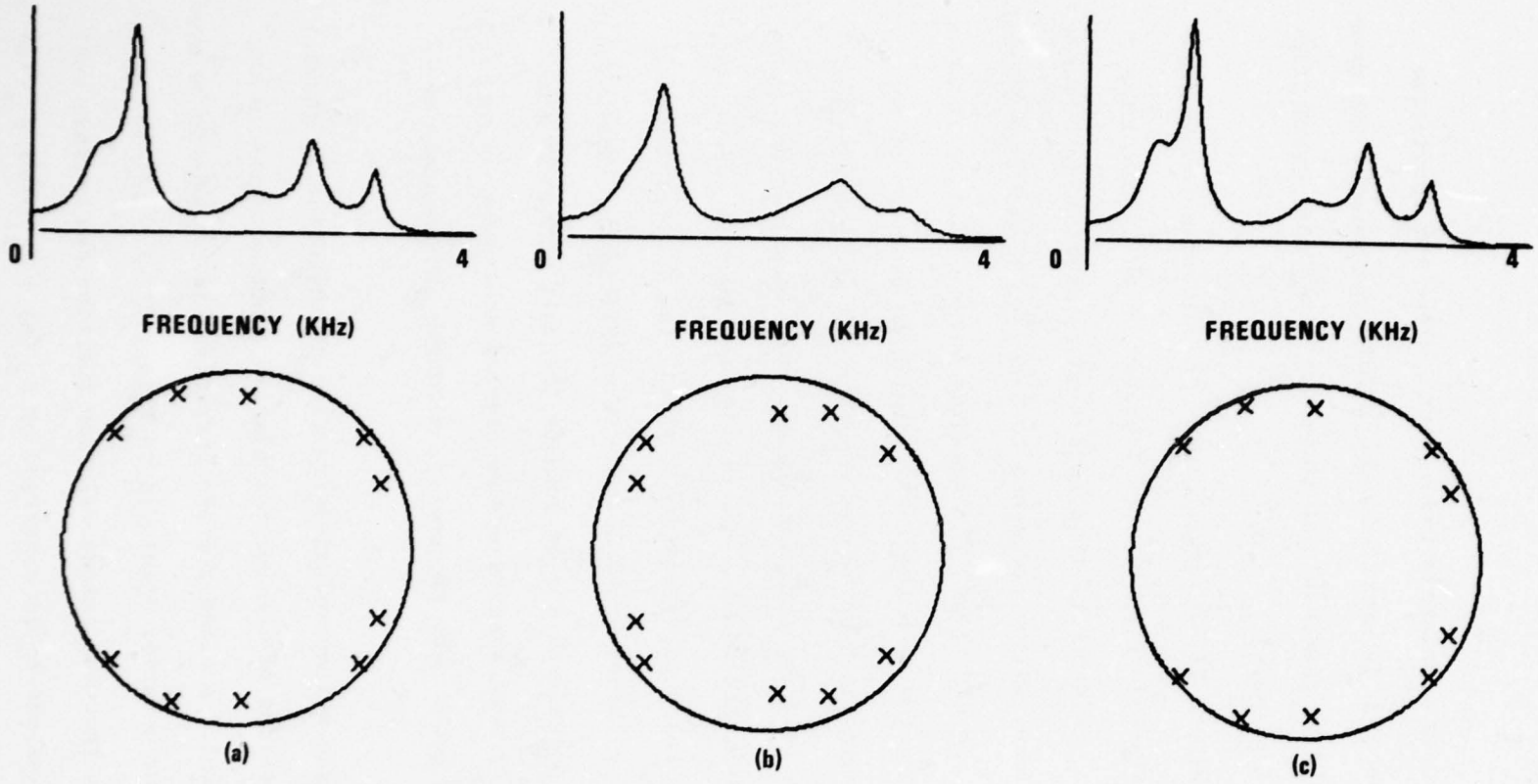


FIGURE III–8 LPC SPECTRA AND POLE LOCATIONS (a) NO NOISE (b) CVSD NOISE (c) CORRECTION APPLIED TO AUTOCORRELATION FUNCTION.

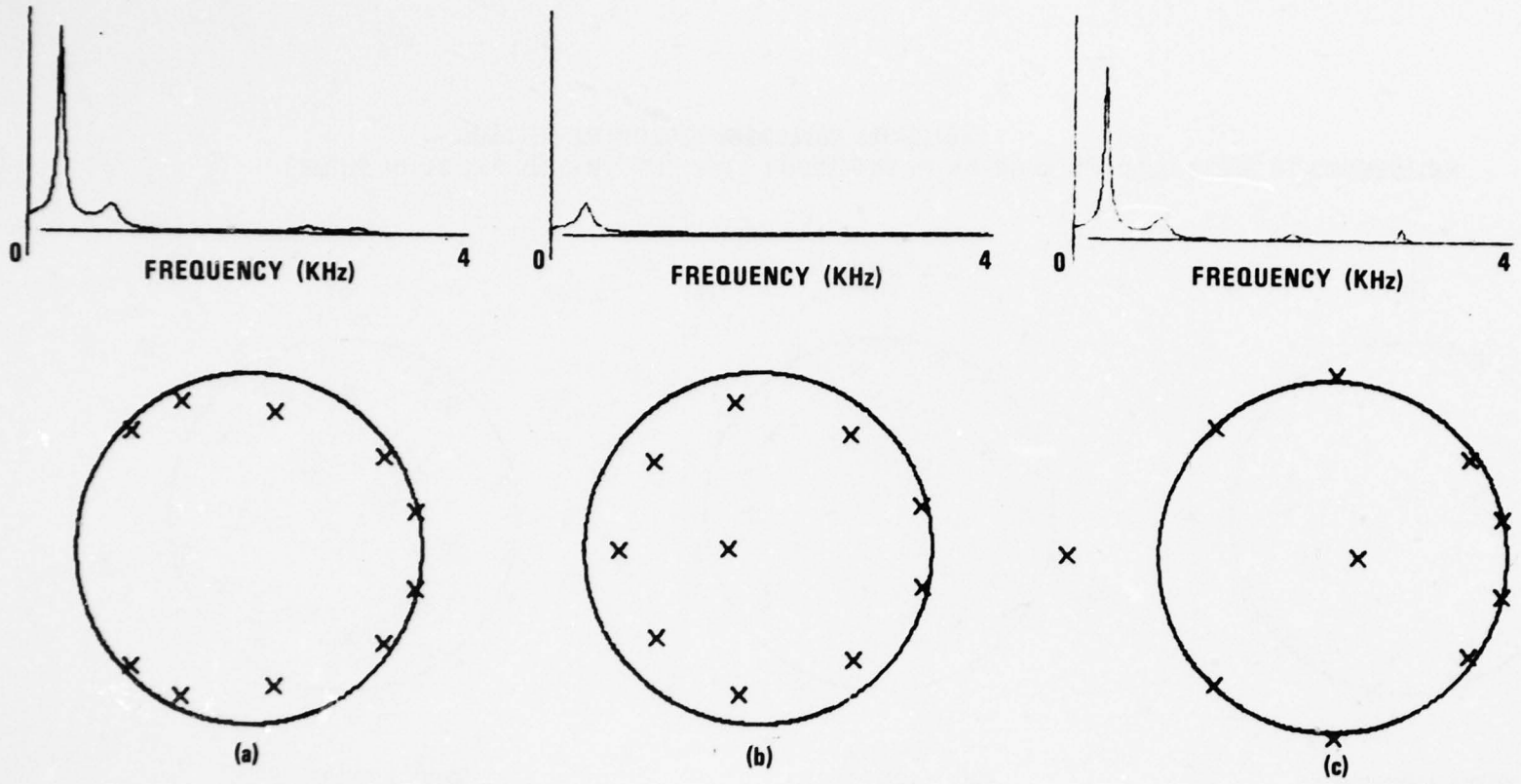
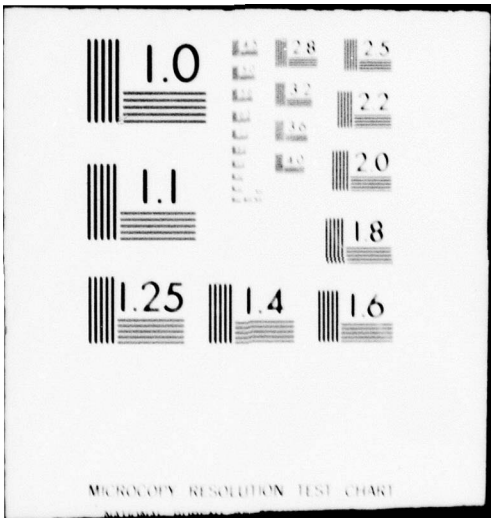


FIGURE III–9 LPC SPECTRA AND POLE LOCATIONS (a) NO NOISE (b) CVSD NOISE (c) CORRECTION APPLIED TO AUTOCORRELATION FUNCTION.

AD-A060 838 GEORGIA INST OF TECH ATLANTA SCHOOL OF ELECTRICAL EN--ETC F/G 17/2
TANDEM INTERCONNECTIONS OF LPC AND CVSD DIGITAL SPEECH CODERS.(U)
NOV 77 T P BARNWELL, R W SCHAFER, A M BUSH DCA100-76-C-0073
UNCLASSIFIED E21-685-77-TB-1 SBIE-AD-E100 092 NL

2 OF 2
AD
A060 838

END
DATE
FILMED
1-79
DDC


1.0
1.1
1.25
1.4
1.6
2.8
3.2
3.6
4.0
2.5
2.2
2.0
1.8
MICROCOPY RESOLUTION TEST CHART

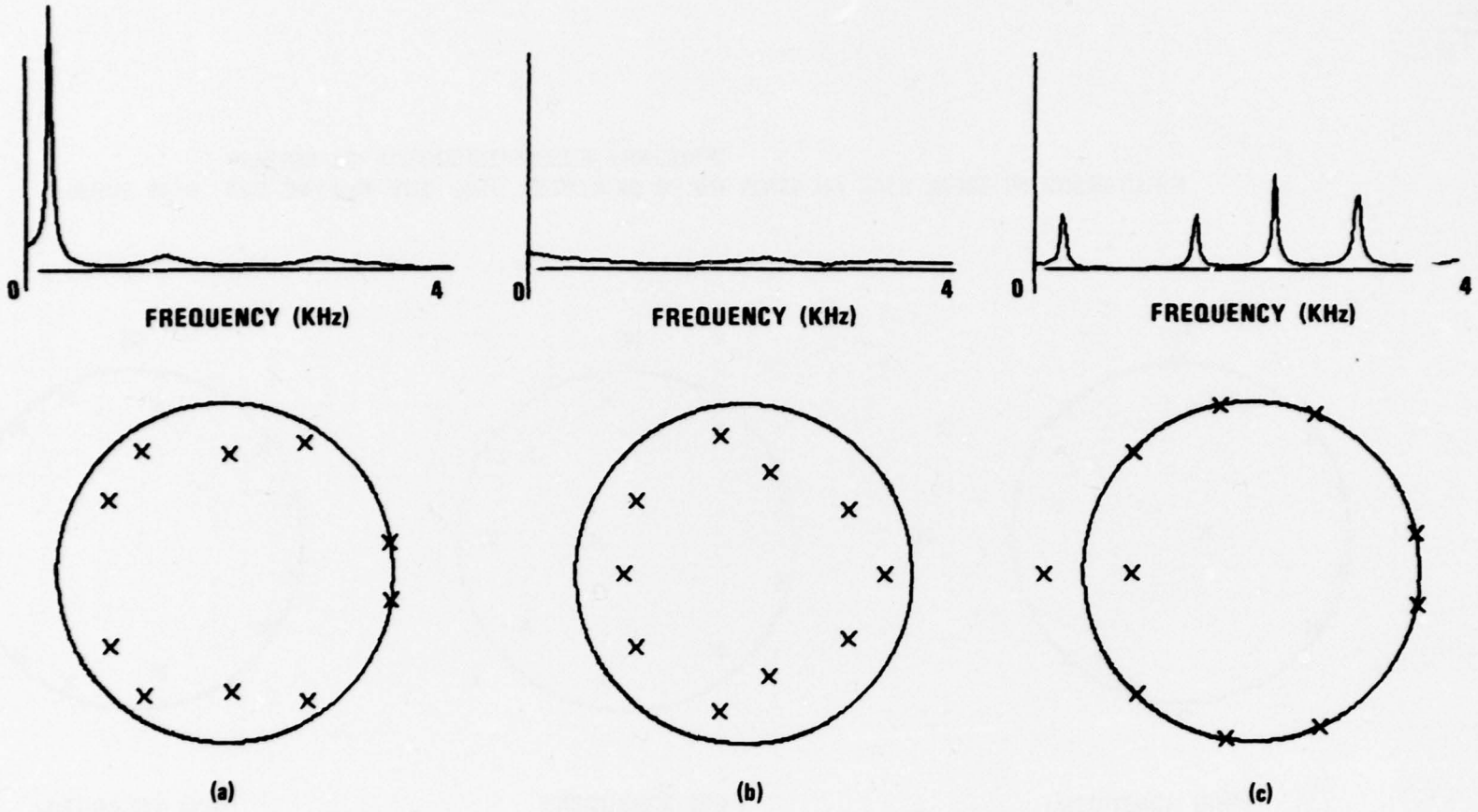


FIGURE III–10 LPC SPECTRA AND POLE LOCATIONS (a) NO NOISE (b) CVSD NOISE (c) CORRECTION APPLIED TO AUTOCORRELATION FUNCTION.

but not by $R_{xx}(m) + R_{cx}(m) + R_{cx}(-m)$. Thus, even though $R_{xx}(m)$ has the correct autocorrelation properties, the cross-correlation terms do not, and, even if small, may cause the LPC analysis to misbehave. It seems, therefore, that attempts to mitigate the effects of additive noise should precede the computation of the short-time autocorrelation function. The next section discusses one approach of this type.

III-4. An Approach to Reducing the Effect of Noise on LPC Analysis of Speech

A standard technique for reducing the effect of noise is to average a number of measurements of the same quantity. If the noise samples are uncorrelated, the desired quantity adds coherently while the noise does not, thereby emphasizing the desired features with respect to the noise. This general principle can also be applied to reducing the effect of noise on the short-time autocorrelation function of voiced speech. Such an approach is described in this section.

A segment of voiced speech as commonly used in short-time analysis, will typically contain several "pitch periods". If the segment is not too long, these pitch periods will be very similar to one another. If noise has been added to the signal, and if the noise is uncorrelated from period-to-period*, then it is reasonable to suppose that averaging together several pitch periods will tend to suppress the noise. This requires that, for voiced speech, the individual pitch periods be isolated and averaged together to obtain a single average pitch period waveform. This is then the input to the autocorrelation computation. This waveform can be viewed as a single period of a perfectly periodic signal, and therefore, the autocorrelation function takes on a particularly attractive form; i.e.

* In the case of severe slope overload noise, this will not be true.

$$R_{\tilde{x}\tilde{x}}(m) = \sum_{n=0}^{N-1} \tilde{x}(n)\tilde{x}((n+m))_{N_p} \qquad \text{(III.8)}$$

where the notation $((n + m))_{N_p}$ means that the indices inside the double sets of parentheses are interpreted modulo N_p. That is, the signal is assumed to be periodic with period N_p. Note that in this case $\tilde{x}(n)$ represents the average pitch period wave for voiced speech. For unvoiced frames, the autocorrelation function is computed using standard techniques.

This approach to computation of the autocorrelation function for voiced speech was studied by Barnwell [17]. The motivation for this initial study was simply to eliminate the need for the data window while maintaining the advantage of guaranteed stability that is inherent in the autocorrelation LPC method. Barnwell found that the performance of the periodic autocorrelation function in LPC analysis compares very favorably with standard autocorrelation LPC techniques and with the Burg method of LPC analysis. [17]

The effect of the averaged pitch period periodic autocorrelation method is illustrated in Figure III-11. Figure III-11a shows the conventional LPC spectrum of a voiced frame without noise and Figure III-11b shows the corresponding spectrum for LPC coded speech. (This is the same case as Figure III-4.)Figures III-11c through III-11h. show the results for the periodic autocorrelation function with averaging over 1, 2, ..., 6 periods, respectively. This sequence of figures displays the important features of the averaging method. First note that averaging two or three periods improves the resolution of the formants at both high and low frequencies. Averaging more than three periods again tends to biur the weaker formants. This is due to the fact that the formants can change appreciably

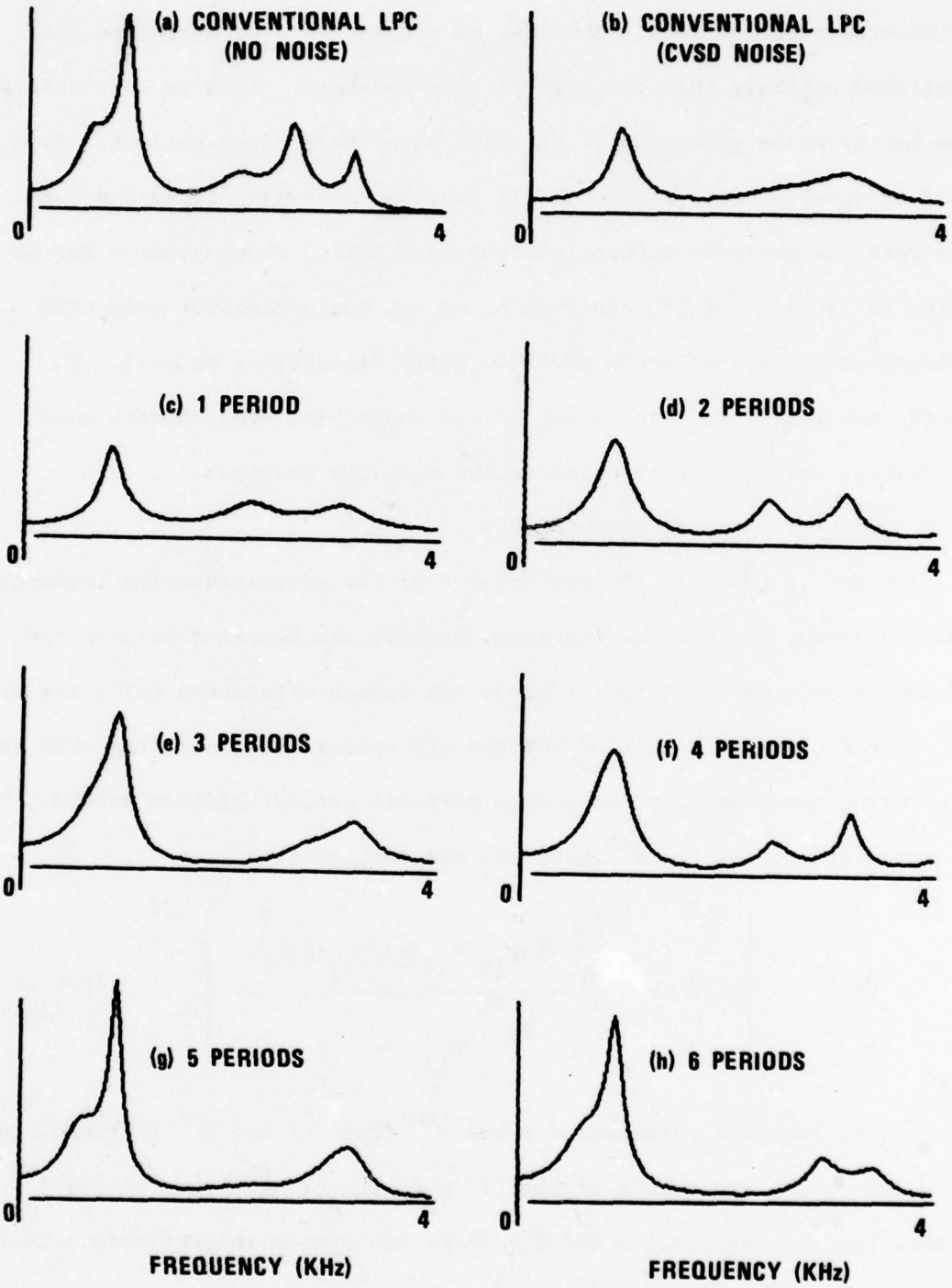


FIGURE III–11 COMPARISON OF LPC SPECTRA. TOP SPECTRA COMPUTED WITHOUT AVERAGING. REMAINING SPECTRA COMPUTED AFTER AVERAGING INDICATED NUMBER OF PITCH PERIODS.

over several pitch periods. This is, of course, more of a problem for low-pitched speakers than for high-pitched speakers. This is unfortunate since better noise rejection is obtained using more pitch periods. Even so, it is apparent from Figure III-11 that the averaging technique combined with the periodic autocorrelation computation shows promise for improving the quality of LPC analysis on speech containing not only CVSD quantization noise, but other additive noise distortions as well. To quantify the benefits, both objective and subjective measurements were made. These results are reported in the next two sections.

III-5. Spectral Distance Measurements

In order to quantify the performance of the noise reduction technique described above, the average spectrum distance was measured between the LPC spectra obtained from the original six speech utterances using the standard autocorrelation LPC method and the LPC spectra obtained from CVSD coded speech using the average pitch period periodic autocorrelation method. The spectrum distance for the i^{th} sentence was defined as

$$D_i = \left\{ \frac{\sum_{J=1}^{96} G_{ij} \sum_{k=0}^{127} \left(10\log_{10}(S_{ij}(k)/S_{ij}(k))\right)^2}{128 \sum_{j=1}^{96} G_{ij}} \right\}^{1/2} \qquad i=1,2,\ldots 6. \quad \text{(III.9)}$$

where G_{ij} is the gain parameter for the j^{th} frame of the i^{th} sentence, and $\tilde{S}_{ij}(k)$ and $S_{ij}(k)$ are values of the LPC spectrum of j^{th} frame of the i^{th} sentence for the test system and the reference system respectively. That is,

$$\tilde{S}_{ij}(k) = \left| \frac{1}{1 + \sum_{m=1}^{10} a_m e^{-j\frac{2\pi mk}{256}}} \right|^2 \qquad \text{(III.10)}$$

where the LPC coefficients $\tilde{a}_m$ were estimated using the average pitch period method, and

$$S_{ij}(k) = \left| \frac{1}{1 + \sum_{m=1}^{10} a_m e^{-j \frac{2\pi mk}{256}}} \right|^2 \qquad \text{(III.11)}$$

where the LPC coefficients a_m were estimated from the original speech signal using the standard autocorrelation LPC analysis method described in Section II-1.2.

The average difference between the LPC spectra computed using the average pitch period method on the original speech and the reference spectra is of interest to provide a point of reference in evaluating the performance of the average pitch period method. Figure III-12 shows the quantity D_i for each of the six sentences of Appendix A as a function of the number of pitch periods involved in the averaging. Note that for one period, which implies no averaging at all, D_i ranges from about 1.7 to 3.5. The dotted curve in Figure III-12 is the average across all six sentences.

$$D = \sum_{i=1}^{6} D_i \qquad \text{(III.12)}$$

as a function of the number of periods included in the average. The results reported in Ref [17] indicate that for this distortion measure, such values imply small distortion. Also note that as the number of periods included in the average increases, the distance also increases. However, the increase is rather modest even out to seven periods.

Figure III-13 shows the results of comparisons when the input to the average pitch period system is CVSD coded speech. In this case the points on the left side represent the distance between the conventional LPC analysis

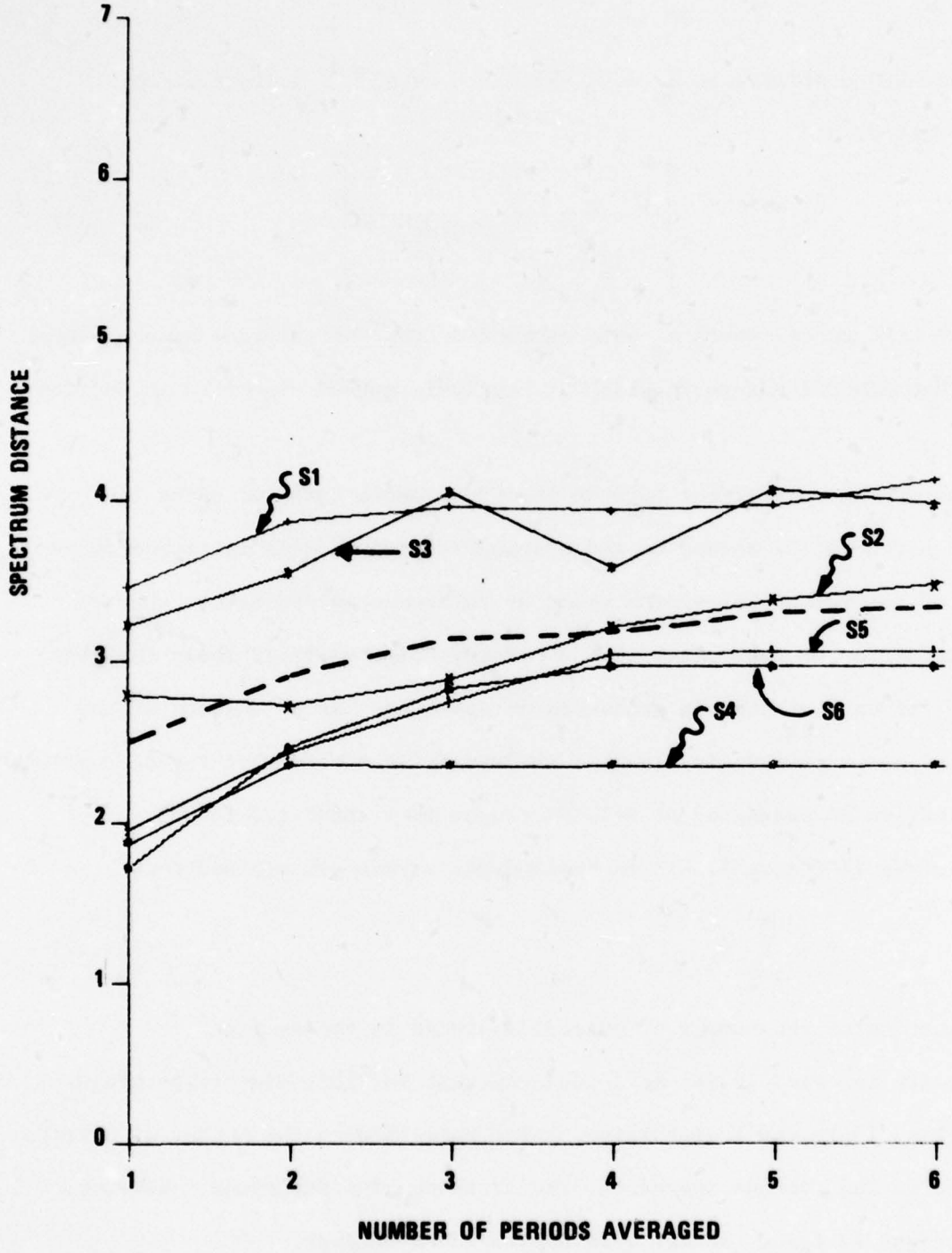


FIGURE III–12 COMPARISON OF AVERAGE PERIOD AUTOCORRELATION METHOD TO STANDARD LPC. DOTTED LINE IS AVERAGE ACROSS ALL SENTENCES.

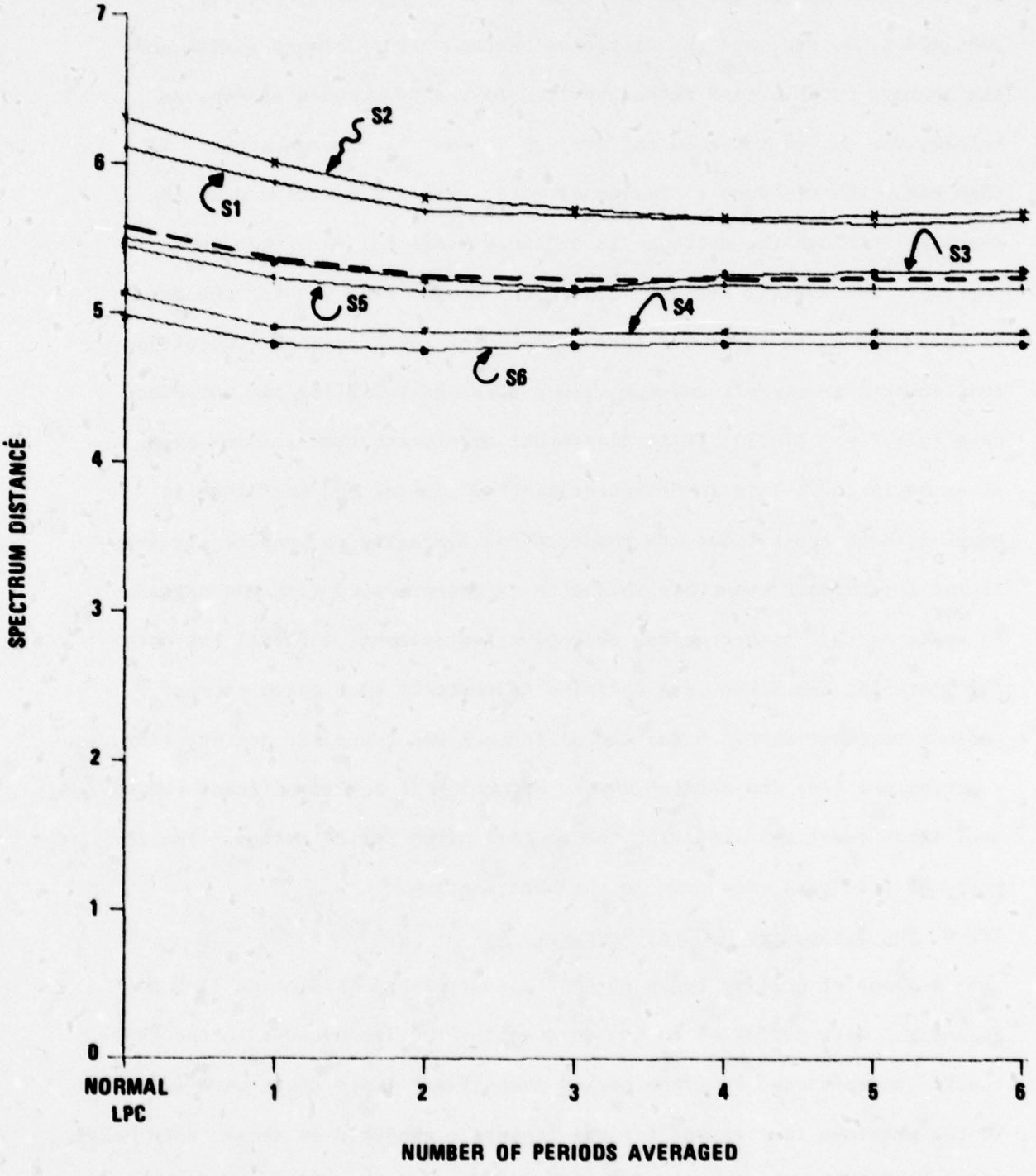


FIGURE III–13 SPECTRUM DISTANCES FOR CVSD INPUT TO NORMAL LPC AND AVERAGE PERIOD METHOD. DOTTED LINE IS AVERAGE ACROSS ALL SENTENCES.

of CVSD coded speech and the reference system. The other points, labelled 2, 3, etc. are the distances between the reference system and the average pitch period method applied to the CVSD coded speech. As before, the dotted curve is the average across all six sentences. In this case, the distance decreases as more periods are included in the average, although the decrease is neglibly small for more than three periods. The overall average distance ranges from 5.6 for the case of no averaging to about 5.2 for averaging of three periods. Recalling that comparable overall average values were about 3.2 for the noiseless case (see Fig. III-12), it is clear that much distortion remains even after averaging. This is not surprising in view of the fact that in general, many repetitions are required for averaging to produce significant improvement even when the noise is uncorrelated with the signal. In spite of this rather modest objective improvement, informal but careful listening tests showed a definite improvement when three or more periods were averaged. Since the difference was certainly not striking, a perceptual test was carried out to determine if any significant perceptual improvement resulted from the average pitch period method. The results of this test are given in the next section.

III-6 The Subjective Quality Test Results

A group of quality tests of the type described in Section II-5 and Appendix C were performed to try to quantify the improvement in the CVSD-to-LPC tandem caused by pitch period averaging. These tests were identical to the previous test except for one feature - subjects in these tests heard and scored the sentences individually, instead of in pairs as before.

In all, three PARM tests were conducted, using 12 subjects each. The systems for the three tests are given in Table III-1, while the test results are given in Table III-2 and Table III-3.

There are a number of points which can be made from the results of these tests. First, it should be noted that, though there is some measureable improvement for multiple period averaging over unaveraged LPC, none of these results are statistically significant at either the .05 or .01 levels. Hence, the differences observed in the careful back-to-back listening tests are not supported by the formal listening tests. Second, it can be seen from these tests that there is a considerable difference in the performance of CVSD and LPC. LPC is consistently 15 points better than CVSD, indicating the considerable difference in the "upper bound" quality of these methods. Last, note that another relatively simple (though not as simple as CVSD) coder was included in the second test, namely a gapped ADPCM [18,19] coder operating at 16 kBPS. This coder has a quality rating of more than 5 points above CVSD, suggesting that this technique, among numerous others, might be used to improve the basic quality of the wideband voice communications system. Further tests of the tandem connection of the gapped ADPCM and LPC systems appear to be of interest.

System	
	PARM #1
1	CVSD
2	CVSD + LPC [autocorrelation]
3	CVSD + LPC [circular with 1 pitch period]
4	LPC [autocorrelation]
	PARM #2
1	16KBPS Gapped ADPCM
2	CVSD + LPC [autocorrelation]
3	CVSD + LPC [circular with 4 pitch periods]
4	LPC [autocorrelation]
	PARM #3
1	CVSD + LPC [circular with 7 pitch periods]
2	CVSD + LPC [autocorrelation]
3	CVSD + LPC [circular with 4 pitch periods]
4	CVSD + LPC [circular with 1 pitch period]

Table III-1 Systems Used In The Three PARM Tests For Testing The CVSD-To-LPC Tandem

PARM #	System HI	1	2	3	4	Low
1	72.0	44.3	34.4	33.3	60.0	23.2
2	71.5	49.6	35.2	35.2	61.6	24.5
3	73.0	40.2	38.0	37.6	39.1	25.2

Table III-2 Means For The Quality Measures For the Subjective Quality Tests

PARM #1

	HI	4	1	2	3	Low
HI		12.1	27.8	37.6	38.7	48.9
4	**		15.7	25.6	26.6	36.8
1	**	**		9.9	10.9	21.1
2	**	**	**		1.1	11.2
3	**	**	**			10.2
Low	**	**	**	**	**	

PARM #2

	HI	4	1	3	2	Low
HI		9.9	21.9	36.3	36.3	47.0
4	**		12.0	26.4	26.4	37.1
1	**	**		14.4	14.4	25.1
3	**	**	**		.0	10.7
2	**	**	**			10.7
Low	**	**	**	**	**	

PARM #3

	HI	1	4	2	3	Low
HI		32.8	34.8	35.0	35.4	47.7
1	**		2.0	2.2	2.6	14.9
4	**			.1	.5	12.9
2	**				.4	12.8
3	**					12.4
Low	**	**	**	**	**	

Table III-3 Statistical Results For The Subjective Quality Tests For The CVSD-To-LPC Tandem

PART IV

The PSP Half Duplex LPC-10 Realization

As part of the contract work reported by this document, a real time half-duplex PSP LPC-10 realization incorporating several of the techniques studied for improved LPC-CVSD and CVSD-LPC tandeming was developed using the DCEC PDP 11/40 graphics system. The resulting realization is a single PSP program, called SUPER, which may be run either as an LPC-10 receiver or an LPC-10 transmitter, but not both. The program could actually be run full-duplex with only minor modifications if a faster version of the PSP processor were available. The program delivered is a modification of a full-duplex program developed by GTE Sylvania for the National Security Agency on the "ADM" Processor. Since the "ADM" version of the PSP processor is faster than those available at DCEC, and since it also has several instructions not available on the PSP, several program modifications had to be made. These included:

1. The replacement of all "ARM" instructions with an appropriate three instruction set for the PSP.
2. The inclusion of code to do actual bit transfers between the two machines [this was not available in the original code].
3. Modification of the system to a half duplex system.

IV-1. Modifications to the Transmitter

The first modification made to the transmitter was to incorporate the circular correlation technique described in Part III into the algorithm. As previously noted [17] this algorithm has several interesting features, including the fact that the covariance and autocorrelation matrices are iden-

tical, and that no window is ever explicitly applied to the data.

The structure of the LPC-10 algorithm made it impossible to realize in the PSP simulations exactly the systems simulated at Georgia Tech. At Georgia Tech, the program was designed to take a fixed number of pitch periods as part of the average, with a limit of 200-300 samples. In the PSP simulation, a limit of 130 samples had to be used due to real time buffering constraints. Hence, on the whole, less of an averaging effect is obtained on the PSP simulations, particularly for low pitched speakers.

The approach taken in modifying the algorithm was to leave the basic algorithm alone, and modify the input data. In a particular frame, if the frame is unvoiced, no action is taken, and the frame is handled as before. If the frame is voiced, however, then an average pitch period is computed over the frame, and this pitch period is placed in the input buffer repeatedly until the buffer is filled. The algorithm then processes this new buffer as before.

Additional modifications were made to the transmitter to allow second order pre-emphasis as discussed in Part II. The first order pre-emphasis filter was replaced with a second order filter with a single complex pole pair located at $r = .8$ and $\theta = .243$ radians.

IV-2. Modifications to the Receiver

The modifications made to the receiver were to incorporate the second order de-emphasis to match the transmitter, and to use an FIR all-pass filter to excite the voiced branch of the synthesizer as discussed in Sec. II-3.3. The realization used was that shown in Figure II-17(a). In this realization, the spectral de-emphasis is done in the normal way, using a minimum phase second order de-emphasis filter to match the pre-emphasis

filter in the transmitter. The phase modification is accomplished by applying the 32 point FIR all-pass approximation filter of Figure II-15 to the voiced segment by utilizing the FIR filter response as in input pulse shape for the vocal tract filter.

IV-3. Running the PSP Simulation

The PSP simulation program uses the front panel switches . The new switch assignments are given in Table IV-1.

Bit	Meaning	
	0	1
7	Run	Hold for Change
4	Receiver	Transmitter
2	ATAL	Circular Correlation
Ø	1st Order Pre-Emphasis	2nd Order Pre-Emphasis

Table IV-1 Switch Settings for the PSP Simulation Program

APPENDIX A

Test Utterances Used In Simulations

The six test utterances used in this study were:

S1. The pipe began to rust while new. (Female speaker)

S2. Add the sum to the product of these three. (Female speaker)

S3. Open the crate but don't break the glass. (Male speaker)

S4. Oak is strong and also gives shade. (Male speaker)

S5. Thieves who rob friends deserve jail. (Male speaker)

S6. Cats and dogs each hate the other. (Male speaker)

These utterances were compiled by the Defense Communication Agency for use in pitch and voicing studies. The speakers represent a large range of pitch characteristics. The sentences are from the 1965 Revised List of Phonetically Balanced Sentences [A1]. The utterances were sampled at 8.0 Hz and quantized to 12 bit linear PCM resolution.

APPENDIX B

INTERPOLATION AND DECIMATION BY A 2:1 RATIO

In simulating the CVSD system, there is a need for changing the sampling rate of speech signals back and forth between 8 kHz and 16 kHz. This appendix summarizes the relevant details of this process, and describes some special characteristics of the computer implementation.

B.1 Sampling Rate Increase (Interpolation) by 2:1

Beginning with samples of an analog waveform at the low rate,

$$x(n) = x_a(nT) \tag{B.1}$$

it is desired to obtain samples at the high rate

$$y(n) = x_a(nT') \tag{B.2}$$

where $T' = T/2$. In the simulation, $1/T = 8$ kHz and $1/T' = 16$ kHz. It is, of course, assumed that no aliasing occured in the initial sampling.

The general approach to obtaining the interpolated signal $y(n)$, is described in [9]. The case of 2:1 increase in sampling rate, is depicted in Figure B.1. First, the sampling rate is increased by 2:1 by filling in a zero sample between each sample of the original signal; i.e.

$$\begin{aligned} v(n) &= x(n/2) && n = 0,\pm 2,\pm 4, \ldots \\ &= 0 && \text{otherwise} \end{aligned} \tag{B.3}$$

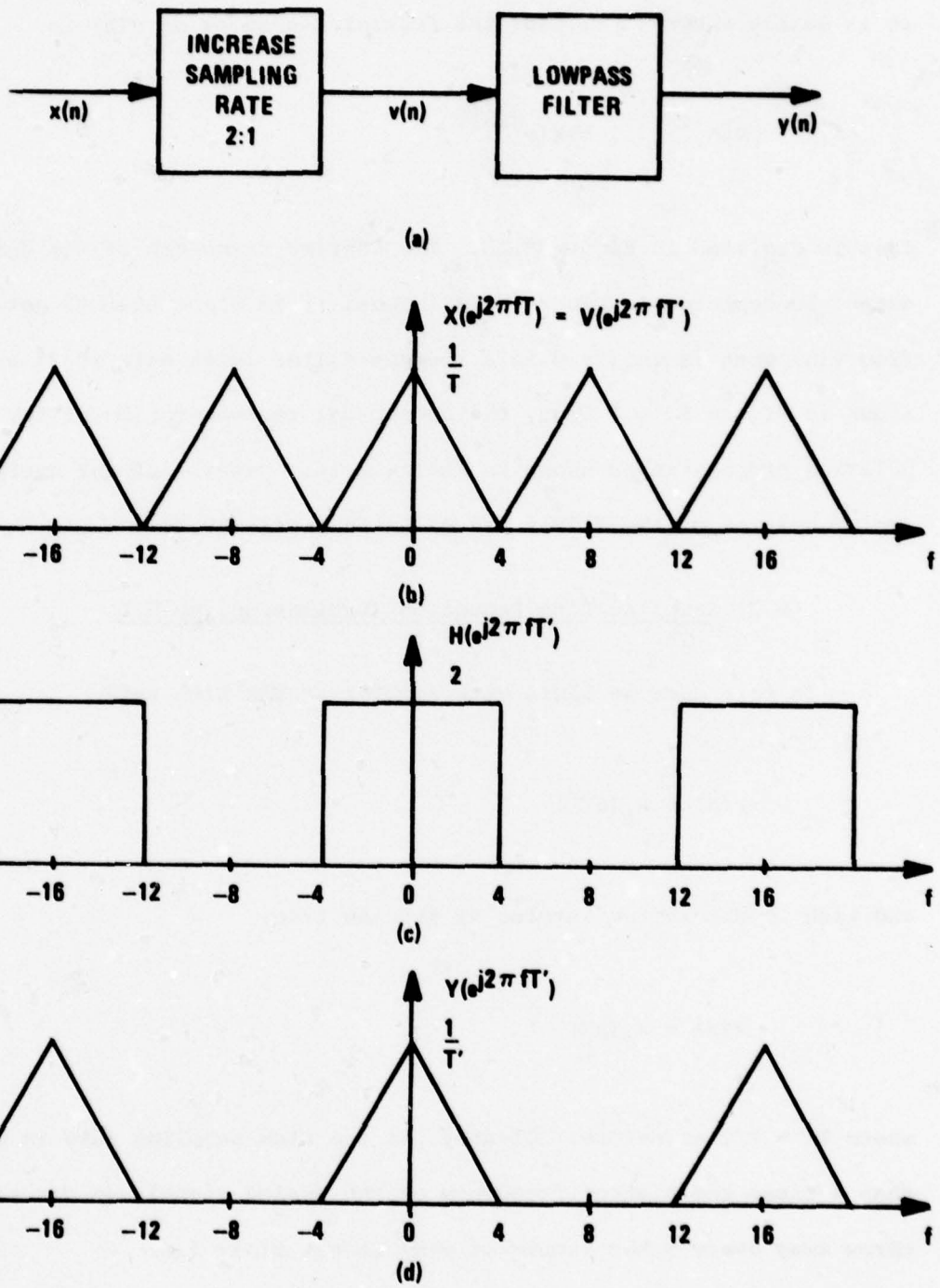


FIGURE B1 SAMPLING RATE INCREASE (INTERPOLATION) BY 2:1. (a) REPRESENTATION OF PROCESS. (b) FOURIER TRANSFORM OF x(n) AND v(n). (c) FREQUENCY RESPONSE OF IDEAL INTERPOLATING FILTER. (d) DESIRED OUTPUT SPECTRUM.

It is easily shown [9], that the Fourier transform of v(n) is

$$V(e^{j2\pi fT'}) = X(e^{j2\pi fT}) \tag{B.4}$$

This is depicted in Figure B.1b. The Fourier transform of the desired output is depicted in Figure B.1d. Thus, it is clear that to get y(n) from v(n) what is required is a lowpass filter (with gain of 2) as shown in Figure B.1c. Thus, the conceptual representation of the interpolation process is as shown in Figure B.1a. Details of the design and implementation of the filter are given in Section B.3.

B.2 Sampling Rate Reduction (Decimation) by 2:1

In this case we begin with samples at the high rate

$$y(n) = x_a(nT') \tag{B.5}$$

and wish to obtain the samples at the low rate,

$$x(n) = x_a(nT) \tag{B.6}$$

where T' = T/2 as before. Clearly, if the high sampling rate is greater than 4 times the highest frequency of the analog signal, we can simply throw away every other sample of y(n) to get x(n); i.e.

$$x(n) = x_a(nT) = x_a(n2T') = y(2n) \tag{B.7}$$

However, in reducing the sampling rate of the CVSD output we must note that the CVSD coding introduces high frequency quantization noise which would be aliased into the speech band. Thus, it is necessary to filter the signal at the high sampling rate before reducing the sampling rate. This process is depicted in Figure B.2a.

Figure B.2b shows the signal (+ noise) at the high rate (with noise at high frequencies). It can be shown [9] that the Fourier transform of the input and desired output are related by

$$X(e^{j2\pi fT}) = \frac{1}{2}[Y(e^{j\pi fT}) + Y(e^{j(\pi fT-\pi)})] \qquad (B.8)$$

(neglecting the effects of noise). Thus, the lowpass filter must cut-off at $\frac{1}{4}$ the high sampling rate and it must have a gain of 1 since the factor of $\frac{1}{2}$ in Equation (B.8) automatically changes the amplitude scale in the proper way.

B.3 Design and Implementation of Lowpass Filters

Both cases require a lowpass filter designed to have a cutoff frequency of 4 kHz when filtering data at a 16 kHz sampling rate. For simulations, it is convenient to use linear phase FIR filters since this permits exact compensation of delays when measuring signal-to-noise ratios.

In the simulations of this report, we used a lowpass filter designed by the Kaiser window method [B1]. The impulse response of the unity gain filter is

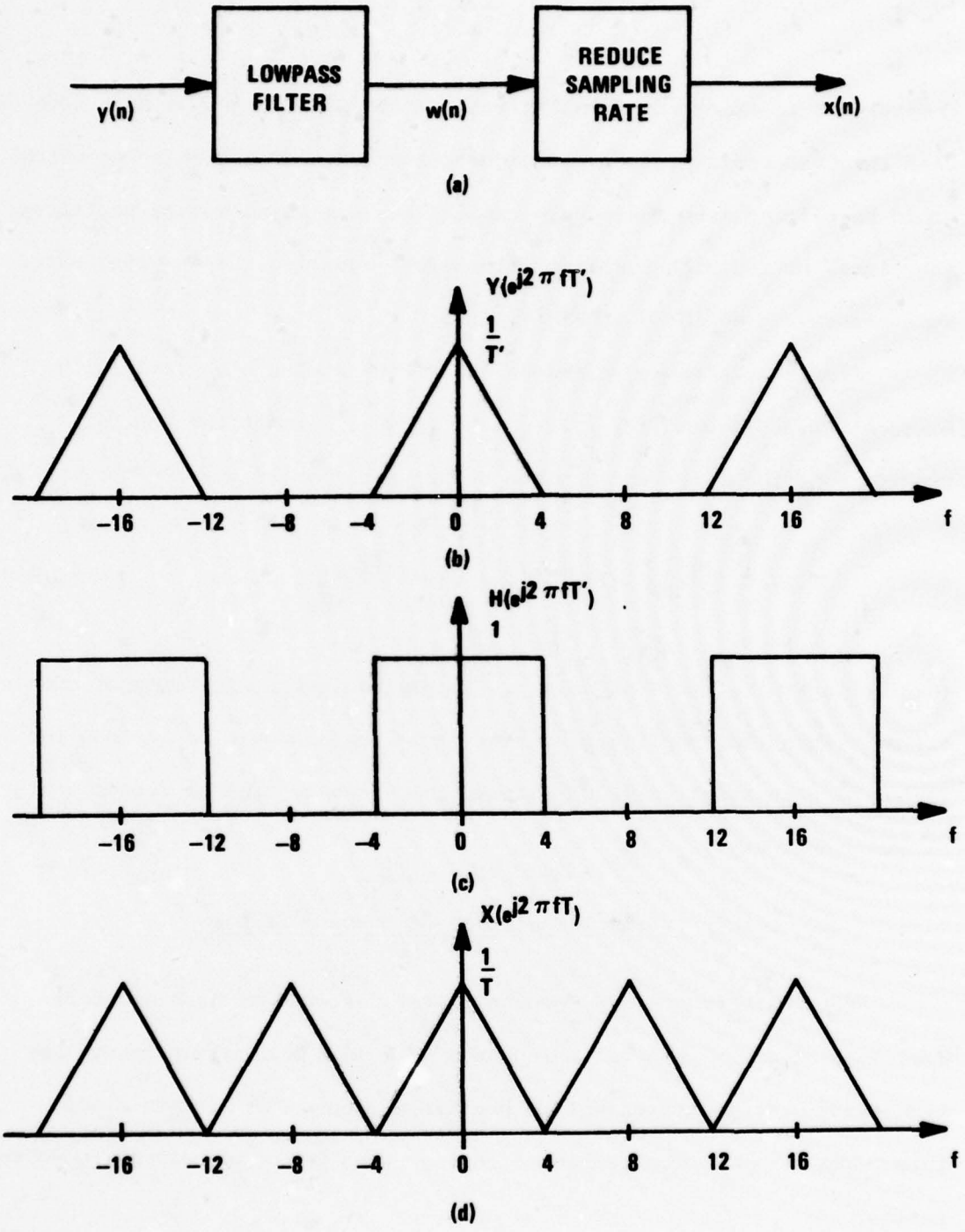


FIGURE B2 SAMPLING RATE REDUCTION (DECEMATION) BY 2:1. (a) REPRESENTATION OF PROCESS. (b) FOURIER TRANSFORM OF INPUT. (c) LOWPASS FILTER REQUIRED TO PREVENT ALIASING. (d) DESIRED OUTPUT SPECTRUM.

$$h(n) = \left[\left(\frac{\sin\frac{\pi}{2}(n-N_d)}{\pi(n-N_d)}\right)\right] w(n) \qquad (B.9)$$

where w(n) is a Kaiser window [B1]. The parameters of the window were set for 50 dB attenuation in the stopband, a nominal cutoff frequency of 4 kHz, a transition width of 1 kHz and a length of 47 samples. The impulse response and frequency response are shown in Figure B.3 and B.4, respectively. In the simulation this filter was implemented with zero phase (N_d=0). For interpolation, the same filter was used with an additional gain of 2.

The particular design discussed above has the following special properties that are advantageous for implementation:

1. It can be seen from Equation (B.9) that

$$h(n) \equiv 0 \qquad n = \pm 2, \pm 4, \pm 6, \ldots \qquad (B.10)$$

   Thus this filter requires only about half the multiplies and adds as a general FIR filter. (This is a consequence of choosing the cutoff exactly at 4 kHz.)

2. The impulse response is exactly symmetric about the central sample ($n=N_d$). Thus half the remaining multiplies can be omitted by adding 2 input samples before multiplication. This property can generally only be expolited in decimation and not in interpolation [9]; however, in the 2:1 case, it can be used to save multiplies in both cases.

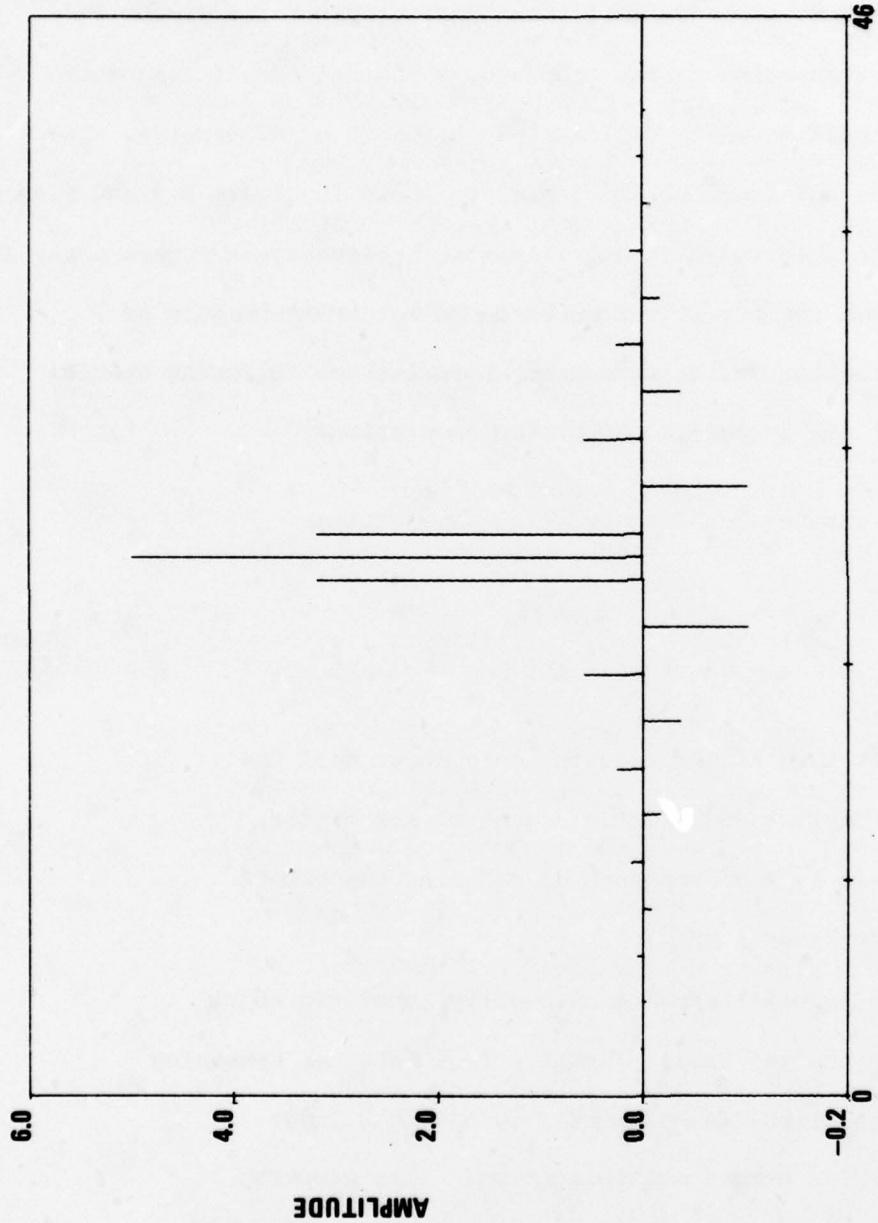


FIGURE B3 IMPULSE RESPONSE OF LOWPASS FILTER USED IN SAMPLING RATE ALTERATION.

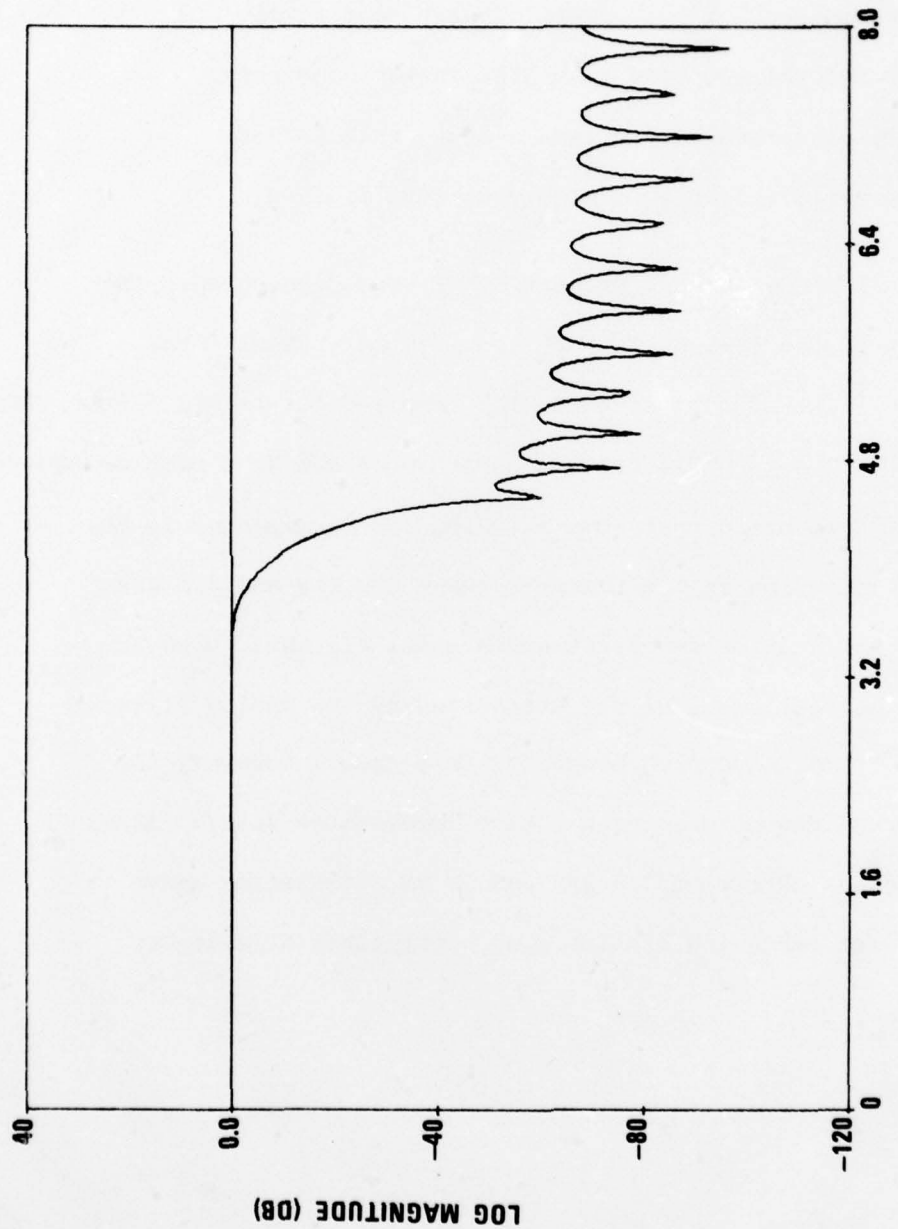


FIGURE B4 FREQUENCY RESPONSE OF LOWPASS FILTER USED IN SAMPLING RATE ALTERATION.

3. In interpolation, every other input sample is zero. This eliminates half the remaining multiplies and adds.
4. In decimation, only every other output sample is need be computed since the intervening ones are discarded. This is a feature that cannot be exploited as well with recursive filters.

Using all these special properties, it can be seen that the implementation of the lowpass filters requires only about $\frac{1}{8}$ the multiplies and $\frac{1}{4}$ the adds as is generally required for an FIR filter. This makes it feasible to use FIR filters in the simulation and the real-time implementation.

It should be noted that some aliasing can be expected in the sampling rate reduction in the region between 3.5 kHz and 4.0 kHz because 50 dB attenuation is not attained until 4.5 kHz. However, this band of frequencies is highly attenuated by the analog filter in the final D-to-A conversion process. As a check, however, the interpolation/decimation process was also implemented with a filter that had a nominal cutoff of 3.5 kHz and 50 dB attenuation above 4.0 kHz. Careful listening disclosed no perceptible differences.

APPENDIX C

Description of Subjective Tests

C-1 Subjective Test Organization

The subjective testing performed here to provide for comparisons among system alternatives was organized along much the same lines as the PARM test [2].

A PARM module is composed of six system configurations: A HIGH ANCHOR, a LOW ANCHOR, and four systems of interest. These systems are presented to the listener in every possible pairing; since there are six systems in the module, this implies 30 possible pairs or 60 total responses for a module. For our test, within a module, five sentences were used, each spoken by a different speaker. The sentences were also presented in all possible pairs, including each sentence paired with itself. Within the pairing constraint cycle, the order of presentation of systems and sentences was chosen at random for the first 30 presentations. The second 30 presentations are a mirror image of the first. This provides a highly balanced context for the presentation of each system configuration. Tables C-1 and C-2 list the keys used for system and sentence organization within a PARM module.

C-2 PARM Data Analysis

Let R(I,J,K,L) represent a raw listener response; where

I denotes the sentence/speaker, $1 \leq I \leq S$

J denotes the presentations of a sentence, $1 \leq J \leq P$

K denotes the system configuration, $1 \leq K \leq C$

L denotes the listeners, $1 \leq L \leq N$

TABLE C - 1

Key for Orderings of Systems

12436541365216532431264523516 4

461532546213423561256314563421

TABLE C - 2

Key for Sentence Ordering

124335413352115324312245235114

4415325442134235512553145534 21

We then compute the averages over responses of each listener to each system:

$$RAVG1(K,L) = \frac{1}{SP} \sum_{I=1}^{S} \sum_{J=1}^{P} R(I,J,K,L)$$

over responses of each listener:

$$RAVG2(L) = \frac{1}{CSP} \sum_{K=1}^{C} \sum_{I=1}^{S} \sum_{J=1}^{P} R(I,J,K,L)$$

over responses to each system:

$$RAVG3(K) = \frac{1}{NSP} \sum_{L=1}^{N} \sum_{J=1}^{S} \sum_{J=1}^{P} R(I,J,K,L)$$

The global or overall average:

$$RAVG\emptyset = \frac{1}{SPCN} \sum_{I-1}^{S} \sum_{J=1}^{P} \sum_{K=1}^{C} \sum_{L=1}^{N} R(I,J,K,L)$$

The ANCHOR averages by listener:

$$RANCH\ (L) = \frac{1}{2}\ (RAVG1(LO,\ L) + RAVG1(HI,L))$$

The overall anchor average:

$$RANCHOR = \frac{1}{N} \sum_{L=1}^{N} RANCH\ (L)$$

We compute the standard errors for the above averages:

$$SE3^2(K) = \frac{1}{(SPN-1)} \sum_{I=1}^{S} \sum_{J=1}^{P} \sum_{L=1}^{N} R^2(I,J,K,L) - \frac{SPN}{SPN-1} RAVG3^2(K)$$

$$SE2^2(L) = \frac{1}{(SPC-1)} \sum_{I=1}^{S} \sum_{J=1}^{P} \sum_{K=1}^{C} R^2(I,J,K,L) - \frac{SPC}{SPC-1} RAVG2^2(L)$$

$$SE\emptyset^2 = \frac{1}{(SPCN-1)} \sum_{I=1}^{S} \sum_{J=1}^{P} \sum_{K=1}^{C} \sum_{L=1}^{N} R^2(I,J,K,L) - \frac{SPCN}{SPCN-1} RAVG\emptyset^2$$

$$SEANC^2(L) = \frac{1}{2SP-1} \sum_{I=1}^{S} \sum_{J=1}^{P} \sum_{K=LO,HI} R^2(I,J,K,L) - \frac{2SP}{2SP-1} RANCH^2(L)$$

$$SEANCH^2 = \frac{1}{2SPN-1} \sum_{I=1}^{S} \sum_{J=1}^{P} \sum_{L=1}^{N} \sum_{K=LO,HI} R^2(I,J,K,L) - \frac{2SPN}{2SPN-1} RANCHOR^2$$

In addition a Newman-Keul test [C1] for the significance of the differences between pairs of system means was carried out for each PARM module. In this test, the system means are ranked, the differences between means are taken, and the studentized range statistic $Q(\alpha,r,f)$, is used to determine the significance of the differences. The statistic is:

$$Q(\alpha,r,f) = \frac{(RAVG3(K) - RAVG3(K'))}{SE1ERROR}$$

with: α = the desired level of significance

r = the number of "steps" between K and K', $2 \leq r \leq C$

f = degrees of freedom of SE1ERROR

= C(SPN-1)

and

$$SE1ERROR^2 = \frac{1}{C} \sum_{K=1}^{C} SE3^2(K)$$

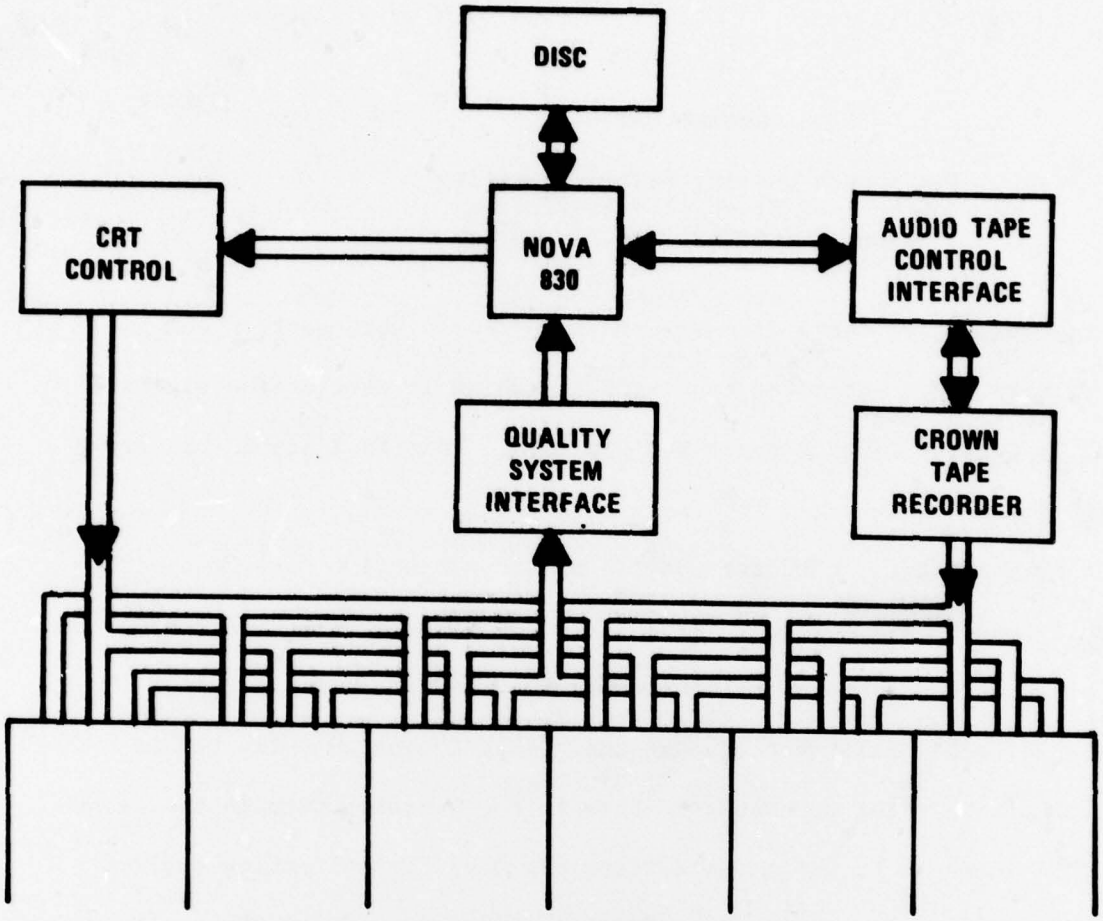
Appendix D

The Speech Quality Testing Facility
School of Electrical Engineering
Georgia Institute of Technology

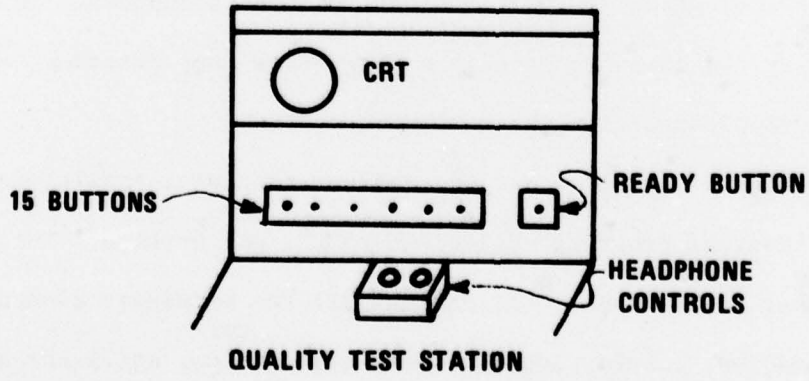
The subjective tests discussed in this report were carried out with the aid of an automated test facility which is part of the Digital Signal Processing Laboratory at Georgia Tech. This facility is briefly described below.

A diagram of the hardware portion of the subjective data acquisition system is shown in Figure D-1. The system consists of six "STATIONS", each of which has an earphone control console, a CRT, and a total of 16 buttons; fifteen "DATA" buttons and one "CONTROL" button. The CRT is used for transmitting alphanumeric data to the subjects through the computer's D/A interfaces, while the buttons are used for collecting subject responses. The audio for the system is supplied by a Crown 800 analog tape recorder which is digitally controlled. In general, 1 kHz tones are placed one track of the analog tape to mark the ends of test sequences. These tones can be detected by the computer through a phase lock loop detector, and are used to accurately position the recorder.

In order to administer the test and collect the data, a multi-task interpretive test control program, called "QUALGOL", was written. The QUALGOL language is summarized in Table D-1, and has all the necessary elements (constants, variables, labels, loop control, arithmetics, etc.) for a simple computer language. Using the QUALGOL language, an experimenter can easily "PROGRAM" a large class of subjective tests on the quality testing facility. A program used for administering some of the tests performed during this study is given in Figure D-2.



SIX STATION QUALITY TESTING FACILITY



QUALITY TEST STATION

FIGURE D–1 AUTOMATED QUALITY TESTING FACILITY.

Table D-1

The QUALGOL Language

Variables

The letters A through Z can be used as variables.

Commands

The following is a list of QUALGOL commands. In this list, V signifies a variable argument and N signifies a constant argument.

Symbol	Function
C(V)	Receive from Crown V = 1 if tone detected V = 0 if no tone detected
C(N)	Send to Crown N = 1 Fast forward N = 2 Stop N = 3 Play N = 4 Record N = 5 Rewind N = 0,6,7 No-op
D(V)	Decrement V by one
DI(N)	Display message N
E	End
G(V)	Get V responses (read buttons) and decrement V to zero
I(V)	Increment V by one
J(V,LABEL) J(@,LABEL)	Jump to LABEL if V=0 Jump to LABEL
K(N)	Light control (off) N = -1 auto off N = 0 red off N = 1 green off
L(N)	Light control (on) N = -1 auto on N = 0 red on N = 1 green on

Symbol	Function
M(N,"Text")	Define message N
P(V)	P nut V
S(V,N)	Let V to N
T	Trace switch
W(N)	Wait N units

```
        M(1,Listen to Sample)
        M(2,Make two responses @ of two digits)
        M(3,Please hurry)
        M(5,              )
        DI(5)L(-1)
        S(I,-30)
        C(3)W(20)C(0)
C1      C(B)J(B,C2)J(@,C1)
C2      C(2)W(1)C(0)
LP      DI(1)S(Z,0)G(Z)
        C(3)W(15)
        C(0)
        W(50)DI(2)S(Z,4)
C8      C(B)J(B,C9)J(@,C8)
C9      C(2)W(1)C(0)
        S(C,-15)
ZX      J(Z,Z2)I(C)W(10)J(C,Z1)J(@,ZX)
X1      J(Z,Z2)DI(3)
        S(C,-10)
Z3      J(Z,Z2)I(C)W(10)J(C,Z2)J(@,Z3)
Z2      K(0)I(I)J(I,EN)J(@,LP)
EN      E
```

Figure D-2 QUALGOL program to control PARM test.

References

1. T. P. Barnwell and A. M. Bush, "A Mini-Computer Based Digital Signal Processing Facility", EASCON '74 Proceedings, October, 1974.

2. W. D. Voiers, "Methods of Predicting User Acceptance of Voice Communication Systems", Final Report, DCA Contract No. DCA-100-74-C-0056, July 15, 1976.

3. N. S. Jayant and A. E. Rosenberg, "The Preference of Slope-Overload to Granularity in the Delta Modulation of Speech", Bell System Tech. J., December 1971, pp. 3117-3125.

4. J. D. Markel and A. H. Gray, Jr., "Linear Prediction of Speech", Springer-Verlag, New York, 1976.

5. J. Makhoul, "Linear Prediction: A Tutorial Review", Proc. IEEE, 63, pp. 561-580, 1975.

6. T. P. Barnwell, J. E. Brown, A. M. Bush, and C. R. Patisaul, "Pitch and Voicing in Speech Digitization", Research Report No. E-21-620-74-BU-1, School of Electrical Engineering, Georgia Institute of Technology, Final Report Submitted to Defence Communications Agency, August, 1974.

7. A. H. Gray, Jr. and J. D. Markel, "Quantization and Bit Allocation in Speech Processing", IEEE Trans Acoustics, Speech and Sig. Proc., Vol. ASSP-24, No. 6, pp. 459-472, December, 1976.

8. A. V. Oppenheim and R. W. Schafer, Digital Signal Processing, Prentice-Hall, Inc., Englewood Cliffs, N.J., 1975.

9. R. W. Schafer and L. R. Rabiner, "A Digital Signal Processing Approach to Interpolation", Proc. IEEE, Vol. 61, No. 6, pp. 692-702, June, 1973.

10. J. L. Flanagan, "Speech Analysis, Synthesis, and Perception", 2nd Edition, Springer-Verlag, New York, 1972.

11. A. V. Oppenheim, "Speech Analysis-Synthesis System Based on Homomorphic Filtering", J. Acoust. Soc. Am., 45, pp. 459-462, 1969.

12. J. N. Holmes, "The Influence of Glottal Waveform on the Naturalness of Speech from a Parallel Formant Synthesizer", IEEE Trans Audio and Electroacoustics, Vol AU-21, No. 3, pp. 298-305, June, 1973.

13. A. E. Rosenberg, "Effects of Glottal Pulse Shape on the Quality of Natural Vowels", J. Acoust. Soc. Am., 49, pp. 583-590, 1971.

14. M. V. Mathews, J. E. Miller, and E. E. David, "Pitch Synchronous Analysis of Voiced Sounds", J. Acoust. Soc. Am., 33, pp. 179-186, 1961.

15. M. R. Sambur and N. S. Jayant, "LPC Analysis/Synthesis from Speech Inputs Containing Quantizing Noise or Additive White Noise", IEEE Trans. Acoust. Speech, and Sig. Proc., Vol ASSP-24, No. 6, pp. 488-493, December, 1976.

16. C. A. McGonegal, L. R. Rabiner, and A. E. Rosenberg, "A Subjective Evaluation of Pitch Detection Methods using LPC Synthesized Speech", IEEE Trans. Acoust., Speech, and Sig. Proc., Vol ASSP-25, No. 3, June, 1977, pp. 221-229.

17. T. P. Barnwell, "Circular Correlation and the LPC", Proc. 1976 IEEE Int. Conf. on Comm., June 1976, Philadelphia, pp. 31-5 - 31-11.

18. T. P. Barnwell and A. M. Bush, "Gapped ADPCM for Speech Digitization", Proc. of National Electronics Conf., Vol. 29, pp. 422-427, Chicago, 1974.

19. T. P. Barnwell, A. M. Bush, J. B. O'Neal, and R. W. Strole, "Adaptive Differential PCM Speech Transmission", RADC-TR-74-177, Final Report, Rome Air Development Center, July, 1974.

A1 "IEEE Recommended Practice for Speech Quality Measurements", IEEE Trans Audio and Electro-acoustics, Vol AU-17, No. 3, pp. 225-246, Spetember, 1969.

B1 J. F. Kaiser, "Non-recursive Digital Filter Design using the I_0-sinh Window Function", Proc. 1974 IEEE Int. Symp. on Circuits and Systems, pp. 123-126, April, 1974.

C1 B. J. Winer, Statistical Principles in Experimental Design, McGraw-Hill Book Co., New York, 1962.